UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

CLASSIFICACIÓ D'OBJECTES MITJANÇANT AGRUPAMENT DE PUNTS D'INTERÈS

Treball Fi de Grau

Presentat a la Facultat

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

per

Jordi Sánchez Escué

En compliment parcial dels requisits per al grau en

ENGINYERIA DE SISTEMES AUDIOVISUALS

Dirigit per

Xavier Giró i Nieto

Carles Ventura Royo

Barcelona, juliol 2014

Abstract

This Bachelor of Science thesis addresses the problem of image classification combining two popular visual representations: points and regions.

Firstly, the study explores bundling interest points with regions. These regions are generated with an initial SLIC partition and using Binary Partition Tree (BPT), considering different scales of resolution in the segmentation.

Secondly, it explores modelling visual classes as a group of points extracted from different images.

Based on Naive-Bayes Nearest Neighbor (NBNN), we are using 1-Nearest Neighbor with SURF descriptor on the *17 Category Flower Dataset* with 1360 images of flowers distributed into 17 classes, 80 images per class.

We have verified that grouping interest points of the same class improves the F1-score a 9.2%. However, bundling interest points into regions using segmentation worsens the F1-score between 1% and 7%, depending on the number of regions in the segmentation.

Resum

Aquest Treball Final de Grau aborda el problema de la classificació d'imatges combinant dues representacions visuals populars: els punts i les regions.

En primer lloc, l'estudi explora l'agrupació de punts d'interès amb les regions. Aquestes regions es generen amb una partició inicial SLIC i s'utilitzen els Arbres de Partició Binària (*Binary Partition Tree*, BPT), considerant diferents escales de resolució en la segmentació.

En segon lloc, s'estudia modelar les classes com a grup de punts d'interès extrets d'imatges diferents.

Basant-nos en el classificador *Naive-Bayes Nearest Neighbor* (NBNN), hem utilitzat el veï més proper amb un descriptor SURF sobre la base de dades *17 Category Flower Dataset*, que conté 1360 imatges de flors distribuïdes en 17 classes, amb 80 imatges per classe.

Hem pogut verificar que el fet d'ajuntar els punts d'interès de les imatges d'una mateixa classe millora la puntuació F1 en un 9,2%. No obstant, l'agrupació de punts d'interès en regions utilitzant una segmentació de la imatge empitjora la puntuació F1 entre l'1% i el 7%, depenent del nombre de regions de la segmentació.

Resumen

En este Trabajo Fin de Grado se aborda el problema de la clasificación de imágenes combinando dos representaciones visuales populares: los puntos y las regiones.

En primer lugar, el estudio explora la agrupación de puntos de interés con las regiones. Estas regiones se generan mediante una partición inicial SLIC y el uso de Árboles de Particiones Binarias (*Binary Partition Tree*, BPT), teniendo en cuenta distintas escalas de resolución en la segmentación.

En segundo lugar, se estudia modelar las clases como grupo de puntos de interés extraídos de imágenes distintas.

Basándonos en el clasificador *Naive-Bayes Nearest Neighbor* (NBNN), hemos utilizado el vecino más cercano con un descriptor SURF sobre la base de datos *17 Category Flower Dataset*, que contiene 1360 imágenes de flores distribuidas en 17 clases, con 80 imágenes por clase.

Hemos podido verificar que el hecho de juntar los puntos de interés de las imágenes de una misma clase mejora la puntuación F1 en un 9,2%. Sin embargo, la agrupación de puntos de interés en regiones utilizando una segmentación de la imagen empeora la puntuación F1 entre el 1% y el 7%, dependiendo del número de regiones de la segmentación.

Agraïments

Abans de res, vull donar les gràcies als professors Xavier Giró i Carles Ventura per la supervisió del projecte. En especial, a en Xavi, per la seva visió general que ens ha permés seguir endavant quan les coses es complicaven massa, i a en Carles, per tot el seu suport a l'hora de tirar endavant el desenvolupament i l'implementació pràctica. També m'agradaria mencionar a la professora Verónica Vilaplana pel suport al llarg de les etapes inicials del projecte.

D'altra banda, m'agradaria agrair el suport per part dels tècnics del Grup de Processament d'Imatge, Albert Gil i Josep Pujal, que sempre han estat disposats a ajudar-me quan han sorgit problemes a l'hora de tirar endavant el desenvolupament del projecte.

Per acabar, agrair a la familia i als amics, a totes aquelles persones que han estat al meu costat durant el projecte i la carrera, que m'han ajudat a seguir fins al final.

Historial i aprovació de revisions

Revisió	Data	Propòsit
0	16/06/2014	Creació del document
1	08/07/2014	Revisió del document
2	09/07/2014	Revisió del document
3	10/07/2014	Aprovació del document

LLISTA DE DISTRIBUCIÓ DE DOCUMENTS

Name	**e-mail**
Jordi Sánchez Escué	jordi.sanchez.escue@alu-etsetb.upc.edu
Xavier Giró i Nieto	xavier.giro@upc.edu
Carles Ventura Royo	carles.ventura@upc.edu

Escrit per		Revisat i aprovat per		Revisat i aprovat per	
Data	08/07/2014	Data	10/07/2014	Data	10/07/2014
Nom	Jordi Sánchez Escué	Nom	Xavier Giró i Nieto	Nom	Carles Ventura Royo
Posició	Autor	Posició	Supervisor	Posició	Supervisor

Taula de continguts

Abstract 1
Resum 2
Resumen 3
Historial i aprovació de revisions 5
Taula de continguts 6
Llista de figures 8
Llistat de taules 9
1. Introducció 10
1.1. Objectius 10
1.2. Requeriments i especificacions 10
1.3. Mètodes i procediments 11
1.4. Pla de treball 11
1.4.1. Work Packages 11
1.4.2. Diagrama de Gantt 12
1.5. Incidències 13
2. Estat de l'art 14
2.1. Ajuntar punts d'interès en una mateixa classe 14
2.1.1. In Defense of Nearest-Neighbor Based Image Classification 14
2.2. Agrupament en farcells de punts 15
2.2.1. Asymmetric Region-to-Image Matching for Comparing Images with Generic Object Categories 16
2.2.2. Altres referències 17
3. Metodologia 18
3.1. Classificador 18
3.2. Extracció de característiques i punts d'interès 19
3.3. Estructura bàsica de classificació 19
3.3.1. Extracció de punts d'interès i característiques 19
3.3.2. Generació de l'ontologia i l'anotació semàntica 20
3.3.3. Entrenament 20
3.3.4. Detecció 20
3.3.5. Avaluació 21

3.4. Ajuntar els punts d'interès en classes 21
3.5. Farcells de punts d'interès en regions 23
3.5.1. Segmentació i jerarquia de particions 23
3.5.2. Detecció 24
3.6. Afegir els punts d'interès en classes i farcells de punts d'interès en regions 25
4. Resultats 26
4.1. Base de dades 26
4.2. Procés d'avaluació 26
4.2.1. Precisió 26
4.2.2. Record 27
4.2.3. Puntuació F1 27
4.3. Resultats de les diferents configuracions 28
4.3.1. Sense ajuntar per classe ni fer farcell de punts 28
4.3.2. Ajuntant per classe, sense farcell de punts 28
4.3.3. Sense ajuntar per classe, amb farcell de punts 29
4.3.3.1. Punts als contorns de les flors 33
4.3.4. Ajuntant els punts d'interès per classe i fent farcells de punts 34
4.3.5. Estudi comparatiu dels resultats obtinguts 35
5. Pressupost 36
6. Conclusions i desenvolupament futur 37
6.1. Conclusions 37
6.2. Desenvolupament futur 37
6.2.1. Altres configuracions a incorporar a l'anàlisi comparatiu 37
6.2.2. Altres classificadors i descriptors 38
6.2.3. Punts aberrants 38
6.2.4. Deixar d'utilitzar els punts d'interès dels contorns 38
6.2.5. Optimitzar la quantitat de regions SLIC a l'hora de generar els BPT 38
6.2.6. Utilitzar segmentacions perfectes per a les imatges d'entrenament 39
6.2.7. Descartar les regions que toquin als marges de la imatge 39
6.2.8. Utilitzar candidats a objecte 39
Bibliografia 40
Annex 42
Vocabulari i glossari 47

Índex de figures

Figura 1: Diagrama de Gantt del Treball Final de Grau 12
Figura 2: Classificació imatge contra classe, de Boiman et al. 15
Figura 3: Classificació amb agrupament de punts d'interès, seguint Kim i Grauman 16
Figura 4: Extracció de punts d'interès, s'ha reduit la quantitat de punts per ser més entendor. 19
Figura 5: Correspondència entre els punts de la imatge a classificar contra els d'entrenament..... 20
Figura 6: Detecció de la classe semàntica segons la imatge més semblant. 21
Figura 7: Correspondència dels punts d'interès de la imatge a classificar contra tots els punts d'interès de totes les imatges d'entrenament d'una mateixa classe. 22
Figura 8: Detecció de la classe semàntica corresponent 22
Figura 9: Arbre de particions binàries (BPT) d'una imatge de la base de dades. 23
Figura 10: Tall de l'arbre BPT anterior, seleccionant les regions utilitzades per la partició N = 3..... 24
Figura 11: Exemple d'una correspondència entre dues regions, marcades en groc en la imatge a classificar i en morat a la imatge d'entrenament..... 25
Figura 12: Comparació dels resultats ajuntant o no els punts d'interès en classes. 29
Figura 13: Comparació fent farcells de punts segons el nombre de regions. 29
Figura 14: Segmentació amb dues regions, una corresponent a la zona de la flor i l'altra al fons..... 30
Figura 15: Segmentació amb dues regions que no es corresponen a la zona flor i zona fons. 31
Figura 16: Segmentació amb tres i sis regions. 31
Figura 17: Correspondència entre dues regions flor, marcades en groc a la imatge a classificar i en morat a la imatge d'entrenament. 32
Figura 18: Correspondència entre la regió flor de la imatge a classificar, marcada en groc, i la regió fons de la imatge d'entrenament, marcada en verd. 33
Figura 19: Punts d'interès propers als contorns de la flor. 33
Figura 20: Comparació de la puntuació F1 fent farcells de punts segons el nombre de regions, incorporant el fet d'ajuntar les regions en classes basat en Boiman et al. 34
Figura 21: Análisis comparatiu dels resultats segons les diferents configuracions utilitzades. 35
Figura 22: Subconjunt de les imatges en què hi ha un bon enquadrament de les flors. 39

Índex de taules

Taula 1: Resultats de la classificació sense ajuntar per classe ni fer farcell de punts. 28

Taula 2: Resultats de la classificació ajuntant els punts d'interès de les imatges de la mateixa classe. 28

Taula 3: Resultats de la classificació fent farcells de punts amb diferent nombre de regions. 30

Taula 4: Resultats de la classificació fent farcells de punts i ajuntant els punts d'interès en classes. 35

Taula 5: Pressupost del projecte. 36

1. Introducció

1.1. Objectius

L'objectiu d'aquest projecte ha estat l'avaluació del potencial d'agrupar punts d'interès (en anglès, *bundling interest points*) d'acord a les regions obtingudes per una segmentació d'una imatge. Per a fer-ho, s'ha realitzat un estudi comparatiu de classificació d'acord amb la utilització dels punts d'interès, tenint en compte aquests punts de forma autònoma o agrupada.

S'ha treballat sobre el conjunt de dades *17 Category Flower Dataset* [1], basat en 1360 imatges, 80 per cadascuna de les 17 classes de flors que conté la base de dades.

1.2. Requeriments i especificacions

Una de les parts més importants d'aquest Treball Final de Grau ha estat el desenvolupament de programari per al Grup de Processament d'Imatge del Departament de Teoria del Senyal i Comunicacions, de manera que podrà ser utilitzat en un futur per la resta de membres del grup i d'altres estudiants.

Els requeriments del projecte venien fixats per:

- Desenvolupament de programari en què s'analitzi les diferents opcions d'agrupament de punts.
- Que el programari en qüestió tingui en compte les diferents estructures de comparació possibles.
- Realització d'un estudi comparatiu amb l'estat de l'art.
- En cas que els resultats siguin adients i hi hagi la possibilitat, presentació a la tasca PlantCLEF Challenge 2014 mitjançant els algoritmes desenvolupats.

Les especificacions del projecte eren les següents:

- Desenvolupat amb llenguatge C++ i Python.
- Desenvolupat mitjançant la plataforma ImagePlus del Grup de Processament d'Imatge (GPI) del departament de Teoria del Senyal i Comunicacions (TSC) de la Universitat Politècnica de Catalunya (UPC).

1.3. Mètodes i procediments

El punt de partida d'aquest projecte és el capítol 4 de la Proposta de Tesis Doctoral d'en Carles Ventura Royo [2]. Moltes tècniques de l'estat de l'art en reconeixement d'objectes es basen en caracteritzar les imatges a partir de les textures al voltant d'uns punts característics detectats automàticament, com és el cas d'aquelles que apliquen descriptors SIFT [3] o SURF [4].

En aquesta proposa de tesi se suggereix que és útil tenir en compte els punts d'interès de forma agrupada en farcells de punts per capturar la configuració espacial entre ells, en lloc de la proposta tradicional d'utilitzar aquests punts d'interès de forma autònoma.

Així doncs, s'ha volgut desenvolupar el programari bàsic que ens ha permès analitzar aquesta hipòtesi i veure futures línies de treball que poden ser aplicades després de la finalització d'aquest projecte.

D'altra banda, s'ha afegit a l'estudi la proposta de Boiman et al [5] respecte a ajuntar totes les imatges d'una mateixa classe a l'hora de fer la classificació, ja que és un dels articles de l'estat de l'art on s'utilitzen els veïns més propers [6] per classificar i on es proposa una millora (el fet d'ajuntar punts d'interès per classe) que ens ha semblat interessant incorporar.

Així doncs, s'ha implementat i avaluat la proposta de Boiman tant en el cas de treballar amb els punts d'interès de forma autònoma (per separat) com agrupada en farcells de punts.

1.4. Pla de treball

La realització del Treball Final de Grau ha anat d'acord amb el pla de treball establert inicialment, a excepció d'algunes modificacions que es van fer a la revisió crítica i alguna incidència comentada a l'apartat 1.5. d'aquesta memòria.

1.4.1. Work Packages

WP 1: Proposta de projecte i pla de treball

WP 2: Anàlisi de l'estat de l'art i de les diferents alternatives possibles

WP 3: Anàlisi del programari disponible a ImagePlus

WP 4: Desenvolupament del programari de classificació

WP 5: Revisió Crítica del projecte

WP 6: Anàlisis de la presentació al PlantCLEF Challenge 2014

WP 7: Avaluació de resultats

WP 8: Redacció i presentació del Treball de Fi de Grau

1.4.2. Diagrama de Gantt

Les modificacions respecte la revisió crítica venen marcades en taronja.

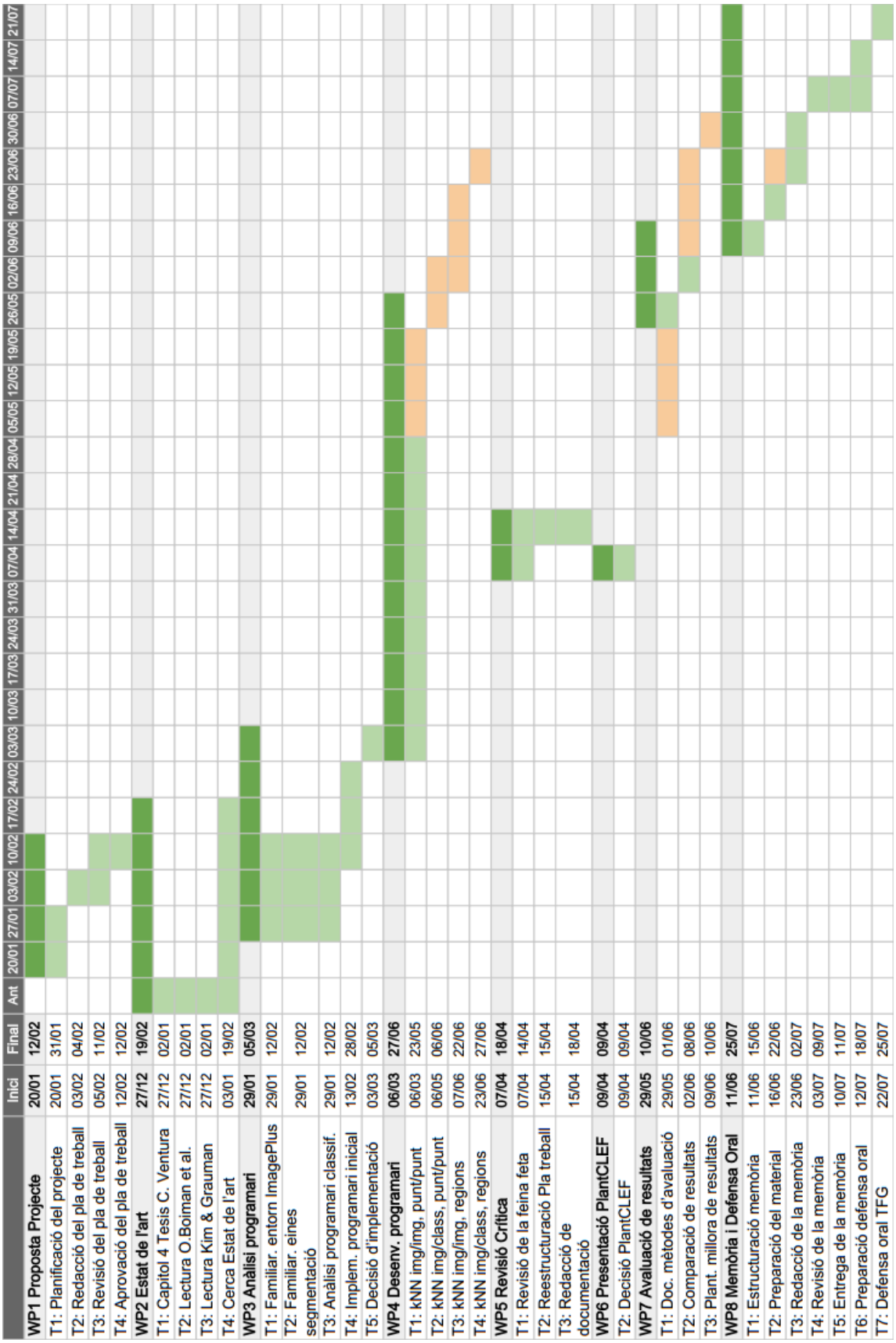


Figura 1: Diagrama de Gantt del Treball Final de Grau

1.5. Incidències

Tal com es pot veure al diagrama de Gantt anterior, la planificació general del projecte ha seguit el que s'havia previst, amb les lleugeres modificacions que es van aplicar al pla de treball durant la revisió crítica del projecte.

Entre d'elles, potser la més significativa, el fet de descartar la participació al PlantCLEF degut a la quantitat de recursos de memòria per a processar aquesta base de dades, fet que comportava una dedicació important en millorar l'eficiència en comptes de centrar-nos en seguir el desenvolupament de les línies bàsiques del projecte, que ja es podien treballar sobre la base de dades *17 Category Flower Dataset*.

A més a més, també s'ha descartat la tasca T5 del WP4, que tractava de valorar el fet de tenir un conjunt d'imatges d'una mateixa classe com a entrada del sistema, ja que ens hem volgut centrar en els elements essencials del projecte, és a dir, el fet d'ajuntar els punts d'interès en classes i l'agrupament d'aquests punts d'interès en farcells de punts segons la seva regió.

S'ha realitzat una feina molt important a nivell de desenvolupament de programari, desenvolupant eines addicionals i preparant una estructura de classificació que fos el més genèrica possible, de manera que es pugui adaptar per a d'altres sistemes de classificació o altres descriptors visuals.

És per aquest motiu que el WP4 s'ha allargat més del previst, però gràcies a la planificació prèvia això no ha afectat directament a la redacció i revisió final de la memòria, que s'han desenvolupat amb normalitat.

2. Estat de l'art

2.1. Ajuntar punts d'interès en una mateixa classe

La primera aproximació que té aquest projecte és el fet d'ajuntar els punts d'interès extrets d'una imatge amb tots els punts d'una mateixa classe. Aquesta proposta ve presentada per Boiman et al a *In Defense of Nearest-Neighbor Based Image Classification* [5].

2.1.1. In Defense of Nearest-Neighbor Based Image Classification

Els classificadors basats en models per la classificació d'imatges requereixen una etapa intensiva d'aprenentatge i entrenament (en anglès, *training*), com per exemple, el cas dels *Support Vector Machines* (SVM).

En canvi, els sistemes de classificació basats en sistemes no paramètrics com el del veí més proper (en anglès, k *Nearest-Neighbor*) amb k = 1 i amb només un descriptor, no necessiten cap temps d'entrenament (només un procés d'indexació de les imatges d'entrenament).

Tot i així, hi ha un gran salt entre el rendiment experimentat pels sistemes entrenats respecte els sistemes basats en veïns més propers. Segons Boiman et al, l'efectivitat d'aquests sistemes no paramètrics ha estat subestimada degut a dos problemes essencials:

(i) la quantificació dels descriptors locals utilitzats, com per exemple la tècnica de Bossa de Descriptors (en anglès, *Bag of Features*, BoF)
(ii) el fet d'utilitzar distàncies imatge contra imatge, en comptes de treballar amb imatge contra classe.

La quantificació ens permet reduir la dimensionalitat en gran mesura, però també provoca una degradació significativa de la discriminació dels descriptors. És per aquest motiu que en el seu treball s'afirma que és innecesari (i perjudicial) fer-ho en el cas de classificadors no paramètrics, ja que no tenen cap fase d'entrenament que compensi aquesta pèrduda d'informació.

Pel que fa a la distància utilitzada, les distàncies imatge contra imatge donen un bon resultat quan les imatges a classificar són semblants a alguna de les imatges d'entrenament de la base de dades, però si no és així els resultats empitjoren considerablement

En canvi, si utilitzem una distància imatge contra classe, podem detectar correctament una nova imatge a classificar que, tot i no ser prou semblant a cap de les imatges de la nostra base de dades, sí que es correspon a la mateixa classe.

Per exemple, en el següent cas, la nova imatge a classificar no es correspon prou bé amb cap de les imatges inferiors (imatges d'entrenament de la base de dades), però si en canvi fem una combinació de totes elles (mitjançant una distància imatge contra classe), millorem en gran mesura els resultats obtinguts.

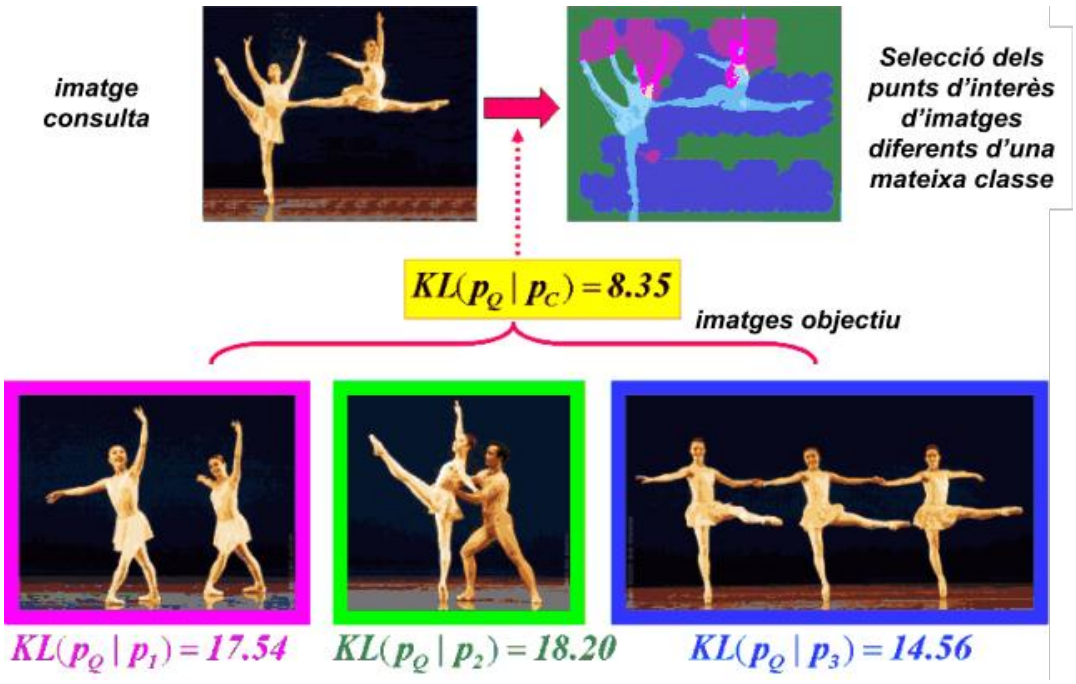


Figura 2: Classificació imatge contra classe, de Boiman et al.

En el treball de Boimann es proposa un sistema anomenat NBNN, Naive-Bayes Nearest-Neighbor, sense quantificació i utilitzant distància imatge contra classe.

Així doncs, en aquest projecte hem volgut analitzar els resultats que es poden obtenir ajuntant els punts d'interès de les imatges d'una mateixa classe a l'hora de generar la classificació final. També s'ha tingut en compte la proposta d'utilitzar l'algoritme de veïns més propers sense cap tipus de quantificació.

2.2. Agrupament en farcells de punts

La segona aproximació d'aquest projecte ve donada per l'agrupament de punts d'interès en farcells de punts segons la seva regió. D'aquesta manera es vol tenir en compte la distribució dels punts en l'espai, tal com proposen Kim i Grauman a *Asymmetric Region-to-Image Matching for Comparing Images with Generic Object Categories* [7].

2.2.1. Asymmetric Region-to-Image Matching for Comparing Images with Generic Object Categories

Kim i Grauman presenten un algoritme de correspondència (en anglès, *matching*) de punts d'interès que aprofita la segmentació en regions d'una imatge. En aquest algoritme, es vol maximitzar la coherència geomètrica i la similitud en aparença. La forma de codificar-ho que es proposa permet que l'algoritme sigui prou flexible per a deformacions no rígides o variacions dins una mateixa classe (intra-classe).

A diferència dels algoritmes convencionals d'imatge contra imatge o dels de regió contra regió, aquest mètode treballa de manera assimètrica, fent correspondre els punts de cada regió d'una imatge segmentada per a una segona imatge sense segmentar.

En aquesta proposta també s'utilitza un algoritme de veïns més propers amb una extracció de punts d'interès densa (mostratge uniforme) a múltiples escales, d'on s'extreuen els descriptors SIFT, generant una graella de punts. Cada posició de la graella, és a dir, cada punt, es representa per dos vectors unidimensionals: un a nivell de columnes i un a nivell de files, en tots dos casos de dalt a l'esquerra a baix a la dreta

D'aquesta manera, com es pot veure en la *Figura 3*, quan es fa la correspondència d'una regió de la imatge segmentada, per cada punt del vector, busquem un conjunt de candidats a la imatge no segmentada. Mitjançant programació dinàmica es minimitza la funció de cost, que té en compte l'aparença i la consistència geomètrica. Aquest procés es genera tant amb el vector d'ordenació en columnes com pel vector d'ordenació en files i es genera la unió de les respectives correspondències.

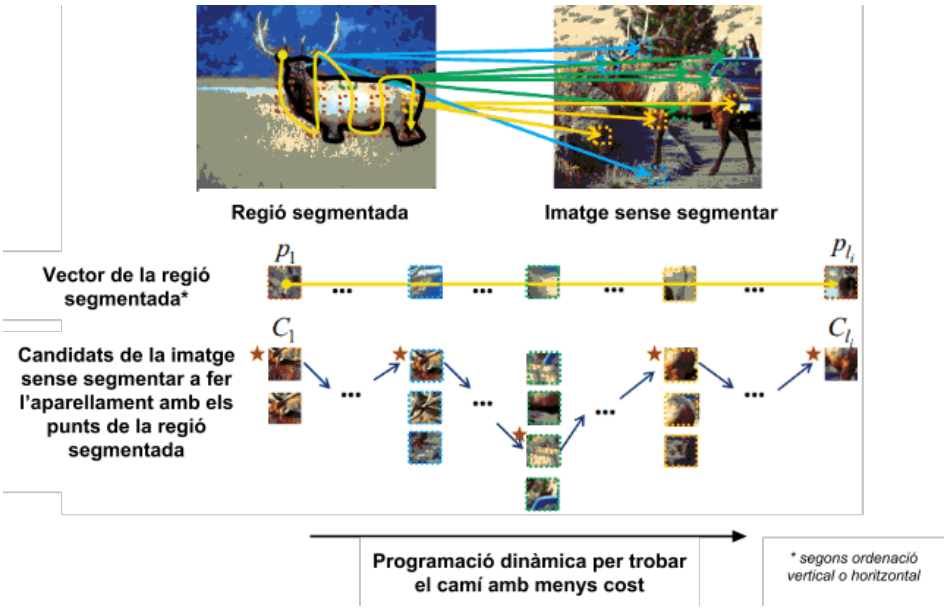


Figura 3: Classificació amb agrupament de punts d'interès, seguint Kim i Grauman

D'aquesta investigació s'ha volgut aprofitar la proposta d'agrupar els punts d'interès en farcells de punts, en especial del fet de fer-ho segons una segmentació de la imatge.

2.2.2. Altres referències

Una altra proposta respecte l'agrupament local ve donada per Wu et al [8], en què s'agrupen els punts d'interès gràcies a dos arguments: (i) els punts agrupats són més discriminatius que un sol punt autònom i (ii) a cada grup es poden fer complir certes restriccions geomètriques simples i robustes de forma eficient.

En el cas de Zhang et al [9], es considera agrupar dos o tres punts d'interès, ja que, afirmen, si se'n combinen masses la repetibilitat de la combinació decreix substancialment. D'aquesta manera, cada agrupació de punts està formada per una punt central i un o dos punts més en un cercle de radi proporcional a l'escala del punt d'interès central.

En aquest cas, la distància entre els dos farcells de punts d'interès té en compte el context espacial, definit per l'orientació i l'escala de les relacions entre els diferents punts d'interès dintre del farcell.

Com que bona part de l'estat de l'art relacionat amb agrupació de punts d'interès no té en compte la informació de contorn a l'hora d'agrupar, en aquest treball volem fer aquesta agrupació tenint en compte una segmentació en regions, de manera que podem comprovar si tenir en compte aquest tipus d'agrupació espacial pot ajudar o no a l'hora de fer la classificació.

3. Metodologia

L'objectiu principal d'aquest projecte era fer un estudi comparatiu que inclogués el fet d'ajuntar els punts d'interés en classes, tal com proposa Boiman, i fer farcells de punts d'interès en regions, seguint les línies principals de Kim i Grauman. És per aquest motiu que la resta de paràmetres de classificació s'han intentat simplificar al màxim, escollint un dels classificadors més senzills i un sol descriptor.

3.1. Classificador

La intenció de la classificació d'imatge és categoritzar una imatge (o fracció d'imatge) en una o diferents classes. Per exemple, volem respondre la següent pregunta: *conté la imatge N un avió?* O bé: *quin tipus de planta conté la imatge M?*

Aquesta classificació es fa mitjançant diferents algoritmes, que es poden diferenciar segons utilitzin classificació supervisada o sense supervisió. La classificació supervisada, utilitzada en aquest projecte, es basa en analitzar un conjunt de dades, anomenat conjunt d'entrenament, que han estat anotades manualment amb la classe o classes corresponents.

Per exemple, en el nostre cas, tenim un conjunt de 1360 imatges anotades cadascuna d'elles amb el tipus de flor que correspon, és a dir, sabent si en la imatge corresponent hi ha una margarita o un gira-sol, per exemple. Això s'anomena anotació semàntica en classes.

D'aquesta manera, donat un nou conjunt d'imatges no anotades, que anomenarem imatges a classificar (o imatges de test), el classificador retorna una o diverses classes, detectant així de quin tipus és més probable que siguin les nostres imatges a classificar, sempre segons el classificador i els paràmetres escollits.

En aquest sentit, s'ha volgut fer ús de la configuració més senzilla de classificació: un classificador basat en els k veïns més propers (kNN), amb k = 1, amb distància euclidiana. També anomenat cerca de proximitat o de similitud, el classificador kNN és el més utilitzat per fer les primeres aproximacions de qualsevol estudi de classificació, quan no tenim informació prèvia sobre la distribució de les dades. Hem utilitzat la distància euclidiana ja que és la recomanada segons OpenCV [10] pel cas de descriptors com el SIFT o el SURF.

Pel cas k = 1, la nostra implementació es basa en, donat un punt d'interès d'una nova imatge a classificar, comparar-lo amb tots els punts de les imatges anotades de la base de dades i retornar el punt més proper. Això es repeteix amb tots els punts de la regió o imatge a classificar. Aquest procediment és explicat amb detall a l'apartat 3.3 i posteriors.

Així doncs, per exemple, donada una nova imatge d'una flor, el classificador ens retorna quina és la imatge més propera a la base de dades i, d'aquesta manera, detecta quina és la classe més probable d'aquesta nova imatge d'entrada.

3.2. Extracció de característiques i punts d'interès

Per classificar una imatge cal que tinguem algunes dades i característiques (en anglès, *features*) per poder-la comparar amb les imatges de la base de dades. Arriba, doncs, el moment de parlar de la detecció de punts d'interès, que serveix per extreure les característiques d'una imatge analitzant els seus punts significatius. Aquests punts es poden extreure de manera densa (en anglès, *dense*), és a dir, mitjançant una quadrícula en què tots els punts tenen la mateixa distància entre ells, o bé de forma dispersa (en anglès, *sparse*), triant punts mitjançant algun algoritme de detecció.

A l'estat de l'art hi ha moltes característiques diferents, en el nostre cas treballem segons l'orientació de diferents filtres de gradient. L'*Speeded Up Robust Features*, SURF [4], és un descriptor, invariant a l'escala i a la rotació, que descriu una distribució de respostes a les *wavelet* de Haar en uns certs punts de les imatges, extraient-ne 64 valors.

S'ha treballat amb el descriptor SURF ja que és molt més fàcil de computar i comparar que d'altres com el SIFT [3], molt freqüent a l'estat de l'art, i aproxima o fins i tot millora els esquemes propostats anteriorment a nivell de repetibilitat, caràcter distintiu i solidesa.

El descriptor SURF va ser presentat juntament amb un detector de punts d'interès dispers, que és el que hem utilitzat per als experiments. Per tant, es deixa per a futurs desenvolupaments l'anàlisi de resultats amb extracció de punts de forma densa.

3.3. Estructura bàsica de classificació

Una vegada conegut el classificador utilitzat, el dels veïns més propers, i el descriptor visual, *Speeded Up Robust Features* (SURF), podem passar a parlar de l'estructura bàsica de classificació utilitzada, que es pot separar en quatre etapes bàsiques.

3.3.1. Extracció de punts d'interès i característiques

La primera etapa de la classificació és l'extracció de punts d'interès i de característiques.

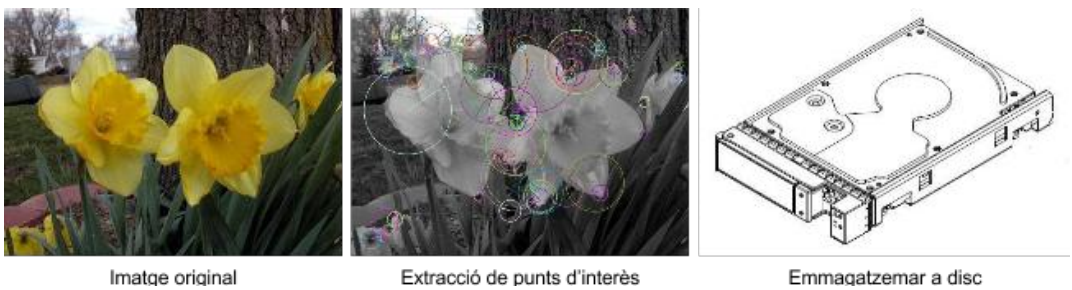


Figura 4: Extracció de punts d'interès, s'ha reduit la quantitat de punts per ser més entendor.

En el nostre cas, hem utilitzat l'eina *bpt_population* ja desenvolupada pel GPI encarregada d'extreure les característiques SURF de les imatges corresponents. Aquesta eina utilitza internament la llibreria OpenCV.

3.3.2. Generació de l'ontologia i l'anotació semàntica

Per a poder fer la classificació necessitem saber de quantes i quines classes disposem. Això és el que en classificació anomenem ontologia. Així doncs, aquesta ontologia conté tots els noms de les classes (que en el nostre cas són els diferents tipus de flor) amb un respectiu identificador numèric per facilitar el desenvolupament de programari.

D'altra banda, abans de poder entrar al procés d'entrenament necessitem tenir una col·lecció d'imatges anotades amb la seva respectiva classe. En el nostre cas, la base de dades *17 Category Flower Dataset* consta. El procés de generació de l'anotació semàntica es basa en processar aquestes imatges i assignar-los la classe corresponent.

3.3.3. Entrenament

El següent pas és entrenar el nostre classificador, en cas que sigui necessari. En el cas dels veïns més propers amb k = 1 i un sol descriptor, aquest pas no és necessari, així que l'únic que ha de fer el sistema és preparar les característiques SURF de les imatges per tal que estiguin organitzades de manera òptima, és a dir, convertir les anotacions en format XML, més pesades i menys eficients, a un format simplificat més ràpid de llegir.

3.3.4. Detecció

La detecció és l'etapa central del procés: s'encarrega de classificar una nova imatge i assignar-li la classe més probable. En aquest pas tenim dos conjunts: els SURF de la nova imatge d'entrada, que estem anomenant imatge a classificar, i els SURF de totes les imatges d'entrenament de la nostra base de dades, ja etiquetades amb la classe.

La configuració bàsica de detecció és la següent:

- Lectura dels punts d'interès i les característiques, emmagatzemades a disc.
- Correspondència dels punts d'interès de la imatge a classificar amb els de totes les imatges d'entrenament, tal com es pot veure a la *Figura 5*.

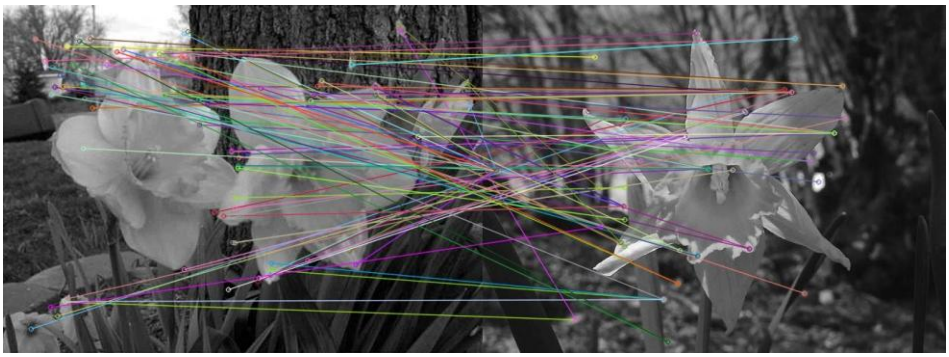


Figura 5: Correspondència entre els punts de la imatge a classificar contra els d'entrenament.

- Selecció de la imatge d'entrenament més semblant a la imatge a classificar.
- En el cas amb k = 1, el classificador decideix que la classe de la imatge d'entrenament més propera és la classe que hauria de correspondre a la nova imatge a classificar, tal com es pot veure a la *Figura 6*.

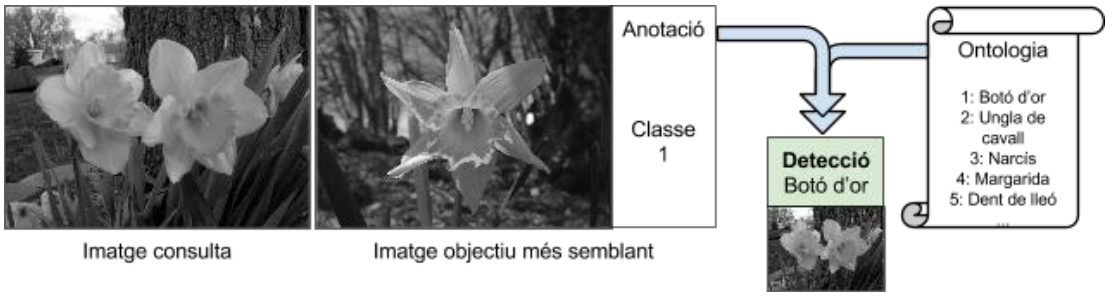


Figura 6: Detecció de la classe semàntica segons la imatge més semblant.

En aquest punt cal explicar com s'efectua la comparació de característiques: en aquest projecte per a fer la comparació s'ha utilitzat una correspondència punt a punt. És a dir: per a cadascun dels punts d'interès de la imatge a classificar, donada una distància (en el nostre cas euclidiana), comparem les característiques amb cadascun dels punts de les imatges d'entrenament de la nostra base de dades i ens quedem amb el més proper i amb la distància corresponent. Aquests punts més propers són els que hem marcat a la imatge de la *Figura 5* que hem vist anteriorment.

Si sumem totes les distàncies de cada punt de la imatge a classificar amb el punt més proper de la imatge d'entrenament, extraiem una distància imatge a classificar contra imatge d'entrenament, que és la que ens permet determinar quina és la imatge d'entrenament més propera de la nostra base de dades.

3.3.5. Avaluació

Al llarg d'aquest projecte hem aprofitat per a desenvolupar una eina d'avaluació, anomenada *Evaluation*. És una eina bàsica que podrà ser ampliada fàcilment, que permet calcular els elements bàsics d'avaluació, que són explicats a l'apartat de resultats.

3.4. Ajuntar els punts d'interès en classes

En l'apartat anterior parlàvem de la configuració bàsica de qualsevol sistema de classificació amb veïns més propers. En aquest i els següents apartats anirem una mica més enllà per explicar els diferents sistemes utilitzats a l'hora de fer l'anàlisi comparatiu.

La primera opció que hem analitzat ha estat el fet d'afegir els punts d'interès segons les classes de les imatges d'entrenament. Aquesta proposta, tal com hem explicat a l'apartat sobre l'estat de l'art, ve presentada per Boiman et al [5] i l'hem volgut incorporar al nostre anàlisi comparatiu per ser una variant d'agrupació de punts d'interès.

En aquesta configuració, les etapes d'extracció de punts d'interès i de característiques, d'entrenament i d'avaluació són les mateixes, però per a fer la detecció cal incorporar algunes variacions. Si fins ara comparàvem la imatge a classificar i la imatge d'entrenament, el que fem ara és, per cada punt de la imatge a classificar, comparem les seves característiques SURF amb tots els punts de totes les imatges d'entrenament d'una mateixa classe, quedant-nos amb el punt més proper de tota la classe.

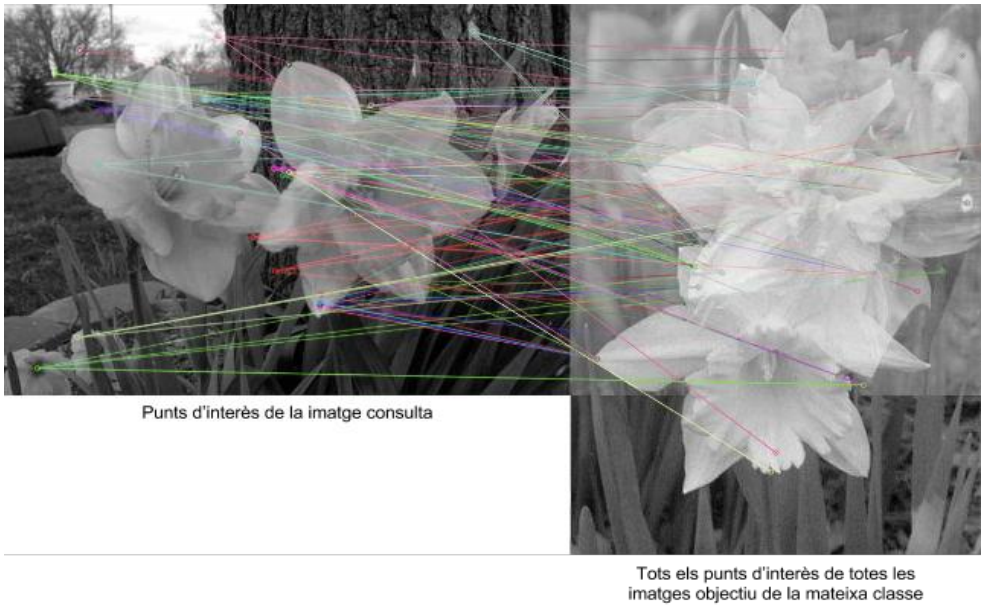


Figura 7: Correspondència dels punts d'interès de la imatge a classificar contra tots els punts d'interès de totes les imatges d'entrenament d'una mateixa classe.

En aquest cas, quan sumem totes les distàncies punt a punt el que tenim és una distància imatge a classificar contra classe d'entrenament i, per tant, directament ens quedem amb la classe que té la distància més curta, tal com es pot veure a la *Figura 8*.

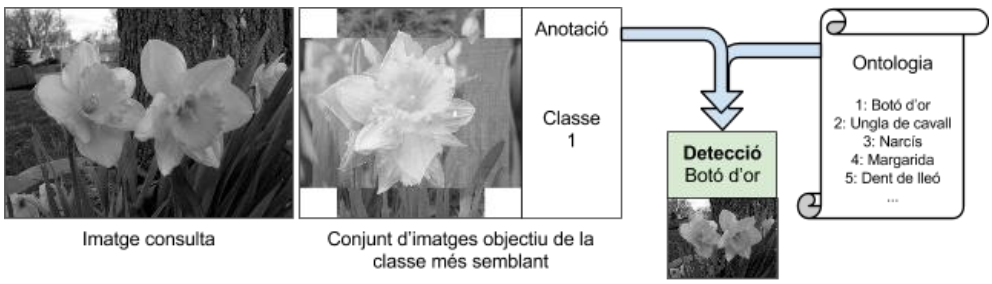


Figura 8: Detecció de la classe semàntica corresponent

3.5. Farcells de punts d'interès en regions

En la tercera configuració de l'experiment, en comptes d'afegir els punts d'interès en classes, el que s'analitza és el fet de fer un farcell (en anglès, *bundle*) dels punts d'interès en regions. Abans hem d'explicar com se segmenten les imatges en regions.

3.5.1. Segmentació i jerarquia de particions

Per tal de segmentar les imatges en regions hem utilitzat una eina ja desenvolupada pel Grup de Processament d'Imatge, anomenada *bpt_creation*, que permet generar el que s'anomenen Arbres de Partició Binària (en anglès, *Binary Partition Tree,* BPT) [11] [12]. Aquests arbres ens permeten, donada una partició inicial, generar una jerarquia binària de particions, és a dir, cada node de l'arbre és una regió que es divideix en dos nodes fills.

Aquesta jerarquia funciona tal com es pot veure a la *Figura 9*, on per cada nivell de la jerarquia, dividim la partició en dues regions més.

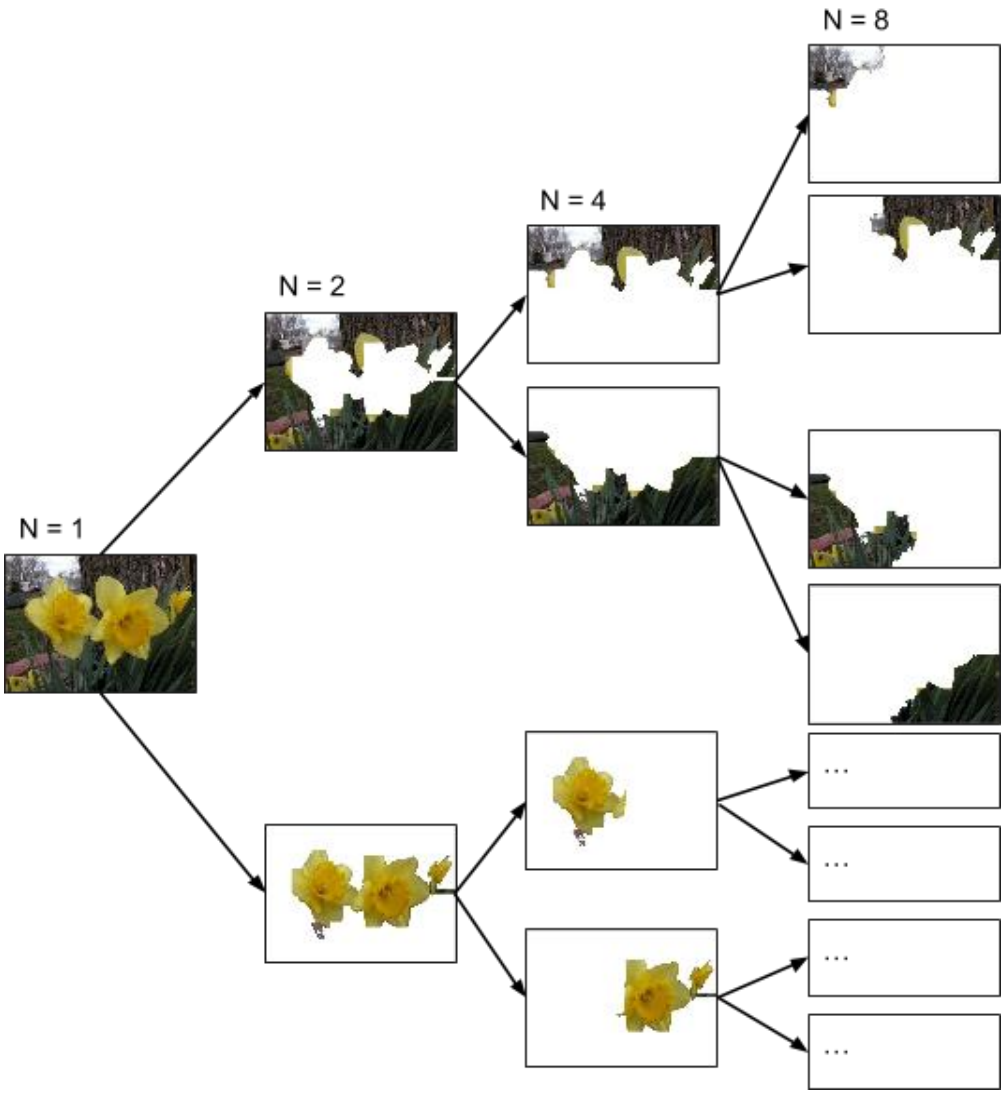


Figura 9: Arbre de particions binàries (BPT) d'una imatge de la base de dades.

De cara a generar la partició inicial, una de les maneres més habituals és la utilització de superpíxels. En el nostre cas hem utilitzat un dels mètodes amb millor rendiment (tant a nivell de resultats com a nivell de velocitat i eficiència), el mètode *Simple Linear Iterative Clustering* (SLIC) [13], gràcies a l'eina *bpt_creation* disponible al Grup de Processament d'Imatge, que realitza la implementació d'Achanta et al [14].

L'únic paràmetre del mètode de càlcul de superpíxels per SLIC és el nombre de regions, que en el nostre cas ha estat de 20. Aquest paràmetre és un dels que es podria variar en línies futures de desenvolupament per optimitzar-lo.

Cal remarcar que, de cara a fer l'estudi comparatiu, no hem utilitzat la jerarquia de particions sencera, sinó només una serie de nodes de l'arbre. Tal com s'explica en l'apartat de resultats, el que fem és tallar l'arbre de manera que ens quedem amb N regions i no amb tot l'arbre sencer. Per exemple, pel cas N = 3, tal com es pot veure a la *Figura 10*.

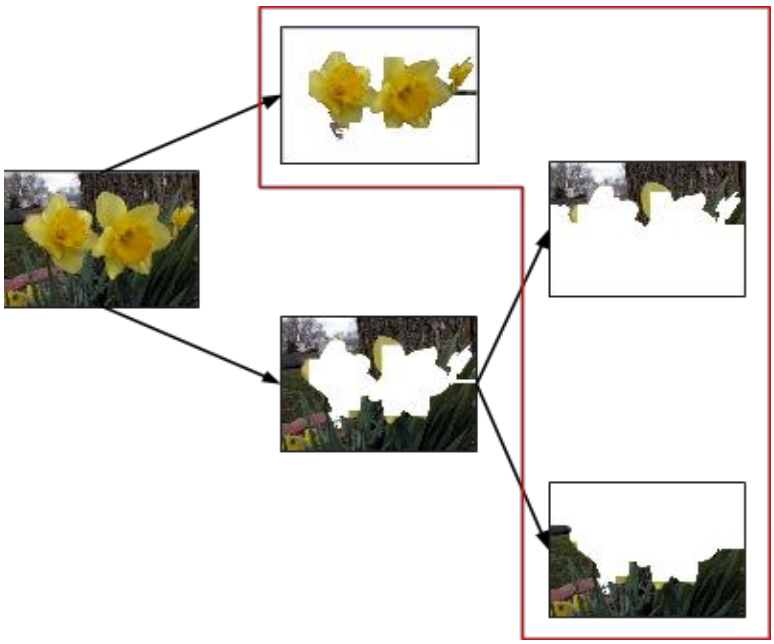


Figura 10: Tall de l'arbre BPT anterior, seleccionant les regions utilitzades per la partició N = 3.

3.5.2. Detecció

Com explicàvem abans, la configuració bàsica es basa en fer una correspondència de punts entre una nova imatge a classificar i cadascuna de les imatges d'entrenament de la nostra base de dades.

En aquest cas el que volem és comprovar el rendiment del classificador afegint una restricció: que tots els punts d'interés d'una regió de la imatge a classificar facin la correspondència amb els punts d'interés, no de tota la imatge d'entrenament, sinó d'una regió concreta d'aquesta imatge. A la *Figura 11* es pot veure un exemple de correspondència entre dues regions.

Així doncs, la distància punt a punt calculada s'acumula per tenir una distància entre les regions, de manera que el classificador pot decidir, donada una regió de la imatge a classificar, quina és la regió més propera de la imatge d'entrenament de la base de dades.

Figura 11: Exemple d'una correspondència entre dues regions, marcades en groc en la imatge a classificar i en morat a la imatge d'entrenament.

Sumant les distàncies de les regions més properes a cadascuna de les regions de la imatge a classificar, es genera una distància que ens permet afirmar quina és la imatge d'entrenament més propera segons aquesta configuració i, utilitzant l'algoritme de veïns més propers, decidir quina classe és detectada per a la nova imatge a classificar.

3.6. Afegir els punts d'interès en classes i farcells de punts d'interès en regions

Havent fet les dues configuracions anteriors, ajuntant els punts d'interés en classes i fent farcells de punts d'interès en regions, una opció final a comprovar és la d'aplicar ambdós sistemes de manera simultània.

Per tant, en aquest últim cas, el que fem és, donada una regió de la imatge a classificar, comparem amb totes les regions de totes les imatges d'una mateixa classe. Així doncs, diferents regions de la imatge a classificar poden correspondre a regions d'imatges d'entrenament diferents, però de la mateixa classe. Sumant les distàncies de les regions més properes, generem una distància "imatge contra classe" que ens permet saber quina és la classe més propera.

4. Resultats

4.1. Base de dades

En aquest projecte s'ha treballat amb la base de dades *17 Category Flower Dataset* [1], una base de dades formada per 17 tipus de plantes, cadascuna d'elles amb 80 imatges anotades. Aquesta base de dades es divideix, com és habitual, en un conjunt d'entrenament (40 imatges per classe), un conjunt de validació (20 imatges per classe) i un conjunt de test (20 imatges per classe).

Els resultats presentats a continuació estan extrets utilitzant 60 imatges a la base de dades (entrenament i validació) i entrant al sistema de classificació les 20 imatges de test, el que hem anomenat fins ara imatges a classificar.

4.2. Procés d'avaluació

Tal com hem comentat anteriorment, s'ha aprofitat aquest projecte per desenvolupar una eina d'avaluació anomenada *Evaluation* a disposició del Grup de Processament d'Imatge. Aquesta eina funciona de la següent manera: (i) donada l'anotació del veritat terreny (en anglès, *ground truth*), és a dir, totes les imatges de la base de dades amb la seva respectiva anotació semàntica i (ii) donats els resultats d'un cert nombre de deteccions, es calculen els paràmetres utilitzats habitualment a l'hora d'avaluar un sistema de classificació: precisió (en anglès, *precision*), record (*recall*) i puntuació F1 (*F1 score*).

Aquests resultats han estat realitzats utilitzant la partició proposada a *17 Category Flower Dataset*, que conté tres configuracions diferents, utilitzant 20 imatges de test cada vegada i amitjanant els resultats (a excepció dels casos agrupant en farcells de N = 12 i N = 15 regions, sense ajuntar en classe, que s'ha calculat amb una sola partició per restriccions de temps de processament i problemes amb el servei de computació).

4.2.1. Precisió

La precisió ens indica quants dels elements predits com a classe N han estat correctament predits.

$$precisió = \frac{predits\ amb\ classe\ N\ correctament}{total\ de\ predits\ com\ a\ classe\ N}$$

No podem avaluar un sistema només amb la seva precisió: imaginem que només hem etiquetat una entrada com a classe N, i l'hem etiquetat correctament. En aquest cas tindríem una precisió del 100%, però potser en comptes d'haver-ne detectat una n'hauríem d'haver detectat deu. És per això que necessitem el càlcul del record.

4.2.2. Record

El record ens indica quants dels elements que hauríem d'haver predit com a classe N segons la veritat terreny realment hem predit correctament.

$$record = \frac{predits\ amb\ classe\ N\ correctament}{total\ d'elements\ classe\ N\ a\ la\ veritat\ terreny}$$

És a dir, seguint amb l'exemple anterior, si hem predit correctament una imatge com a classe N, però n'hauríem d'haver predit deu, tindríem un record del 10%.

4.2.3. Puntuació F1

En tenir dues mesures d'avaluació complementàries apareix un problema essencial: com podem comparar diferents sistemes d'avaluació? Per exemple, es pot donar el cas que un sistema tingui millor precisió i l'altra tingui millor record.

Evidentment, el sistema a triar dependrà segons l'aplicació concreta que estiguem desenvolupant, és a dir, si ens interessa més tenir una bona precisió o un bon record. Tot i així, és bo tenir una mesura única que ens permeti comparar dos sistemes. Aquesta mesura normalment és la puntuació F, que genera el resultat mitjançant una combinació de la precisió i el record:

$$F_\beta = (1 + \beta^2) \cdot \frac{precisió \cdot record}{\beta^2 \cdot precisió + record}$$

El paràmetre β ens permet assignar més importancia a la precisió o al record segons les necessitats de la nostra aplicació. El cas més habitual és per $\beta = 1$, és a dir, aquell en què assignem la mateixa importància a ambdues mesures. En aquest cas, la puntuació F_1 queda de la següent manera:

$$F_1 = 2 \cdot \frac{precisió \cdot record}{precisió + record}$$

I és la puntuació que hem utilitzat en el nostre cas i la que es pot trobar a l'apartat de resultats.

4.3. Resultats de les diferents configuracions

4.3.1. Sense ajuntar per classe ni fer farcell de punts

Els primers resultats venen fixats per la primera configuració del detector utilitzat, és a dir, la més senzilla: donats els punts d'interès i els descriptors SURF de totes les imatges, fer una correspondència punt a punt i quedar-nos amb la millor imatge. Per tant, en aquesta configuració no tenim en compte ni regions ni classes.

En aquest cas els resultats han estat els següents:

Precisió	0,591769
Record	0,381372
Puntuació F1	0,463660

Taula 1: Resultats de la classificació sense ajuntar per classe ni fer farcell de punts.

Aquests resultats ens serviran de referència per a comparar-los amb la resta de configuracions.

4.3.2. Ajuntant per classe, sense farcell de punts.

Un cop analitzats els resultats de la versió bàsica del detector, passem a provar el fet d'ajuntar els punts d'interès per classe, tal com hem explicat al punt 3.4, a la pàgina 21. En aquest cas els resultats han sigut els següents:

Precisió	0,779526
Record	0,432353
Puntuació F1	0,555967

Taula 2: Resultats de la classificació ajuntant els punts d'interès de les imatges de la mateixa classe.

Així doncs, tal com podem comprovar a la *Figura 12*, el fet d'ajuntar els punts d'interès de totes les imatges d'una mateixa classe abans de realitzar la detecció ens permet millorar els resultats de la puntuació F1 de forma substancial, passant d'un 46,4% a un 55,6%.

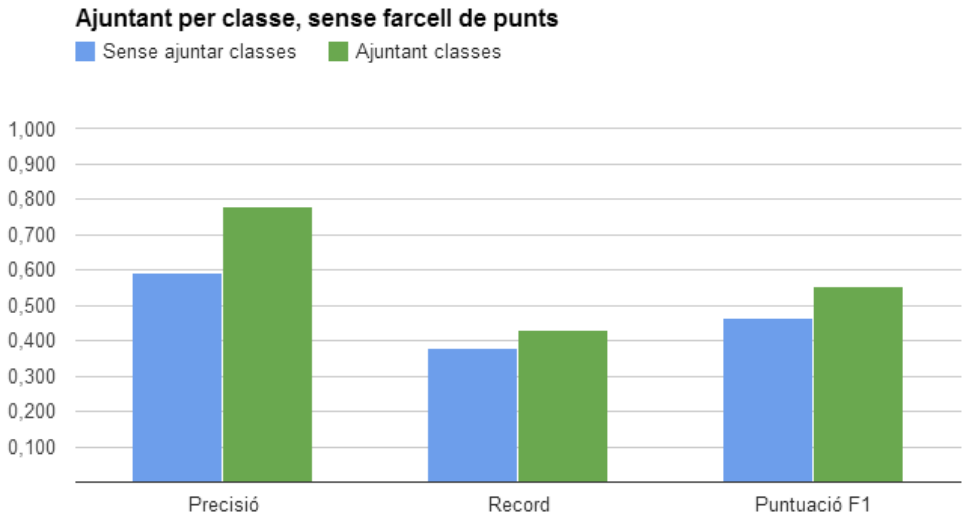


Figura 12: Comparació dels resultats ajuntant o no els punts d'interès en classes.

4.3.3. Sense ajuntar per classe, amb farcell de punts

En aquest punt ens oblidem d'ajuntar els punts en classes i passem a mirar la proposta explicada en el punt 3.5 a la pàgina 23, és a dir, fer un farcell amb els punts d'una mateixa regió. Com ja hem comentat a l'apartat 3.5.1, per a fer la segmentació en regió hem utilitzat BPT amb una partició inicial de superpixels generada amb 20 regions SLIC.

En aquest cas hem analitzat els resultats segons diferent nombre de regions a les imatges, obtenint els resultats següents, que comparem amb el cas bàsic, indicat per N = 1, és a dir, una sola regió (que correspon a no fer els farcells de punts):

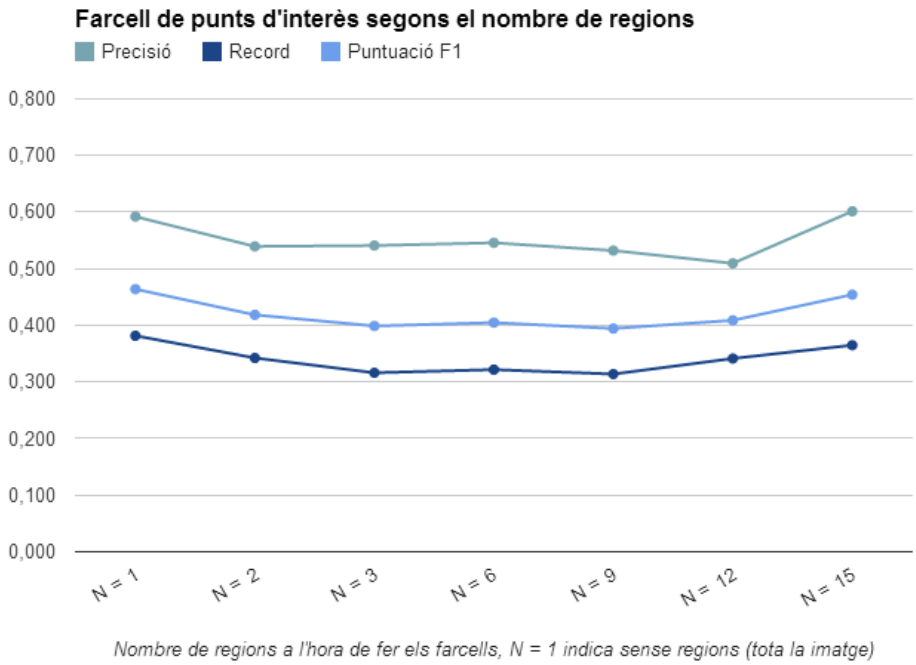


Figura 13: Comparació fent farcells de punts segons el nombre de regions.

Com podem veure a la *Figura 13* i a la *Taula 3*, els resultats no són els que esperàvem: per a qualsevol nombre de regions, la puntuació F1 és inferior a la mesura que trobàvem en el cas sense farcell de punts.

	N = 1	N = 2	N = 3	N = 6	N = 9	N = 12	N = 15
Precisió	0,592	0,539	0,541	0,545	0,532	0,509	0,601
Record	0,381	0,342	0,316	0,322	0,314	0,341	0,365
Punt. F1	0,464	0,418	0,399	0,405	0,394	0,409	0,454

Taula 3: Resultats de la classificació fent farcells de punts amb diferent nombre de regions

En les properes pàgines explicarem el procés que s'ha dut a terme i veurem els possibles motius d'aquests resultats. Per una banda, l'anàlisi inicial d'aquesta configuració l'hem fet amb dues particions ja que, intuitivament i si les imatges són prou contrastades, les dues regions haurien de correspondre amb la part corresponent a la flor i la part corresponent al fons. Un exemple més o menys bo és el que es pot veure a la *Figura 14*.

Figura 14: Segmentació amb dues regions, una corresponent a la zona de la flor i l'altra al fons.

Tot i així, això no sempre és així, ja que en alguns casos les regions generades de l'arbre BPT tallat a dues regions no es corresponen amb el que podria ser intuitiu per a una flor qualsevol, tal com es pot veure a la *Figura 15*.

Figura 15: Segmentació amb dues regions que no es corresponen a la zona flor i zona fons.

Això es pot solucionar en alguns dels casos treballant amb un nombre més elevat de regions, de tal manera que una part de les regions es corresponen al fons i una part a la flor, tal com es pot veure a la *Figura 16*.

Figura 16: Segmentació amb tres i sis regions.

Així doncs, intuitivament es pot esperar que, si fem els farcells de punts per regions, les regions de la imatge a classificar que corresponguin a les zones on hi ha flor faran la correspondència amb les regions de les imatges d'entrenament que també contenen flors, i de la mateixa manera amb les imatges del fons. A la *Figura 17* se'n pot veure un exemple, on la regió "flor" de la imatge a classificar (pintada en groc) correspon amb la regió flor de la imatge d'entrenament (pintada en morat).

Figura 17: Correspondència entre dues regions flor, marcades en groc a la imatge a classificar i en morat a la imatge d'entrenament.

Tot i així, després d'analitzar per què no estaven funcionant correctament els resultats, es pot veure com en una bona part dels casos la correspondència de menor distància es fa entre una regió flor i una regió fons, o viceversa.

A la *Figura 18* se'n pot veure un exemple, amb la regió flor de la imatge a classificar (marcada en groc) fent la millor correspondència amb la regió fons de la imatge d'entrenament (marcada en verd).

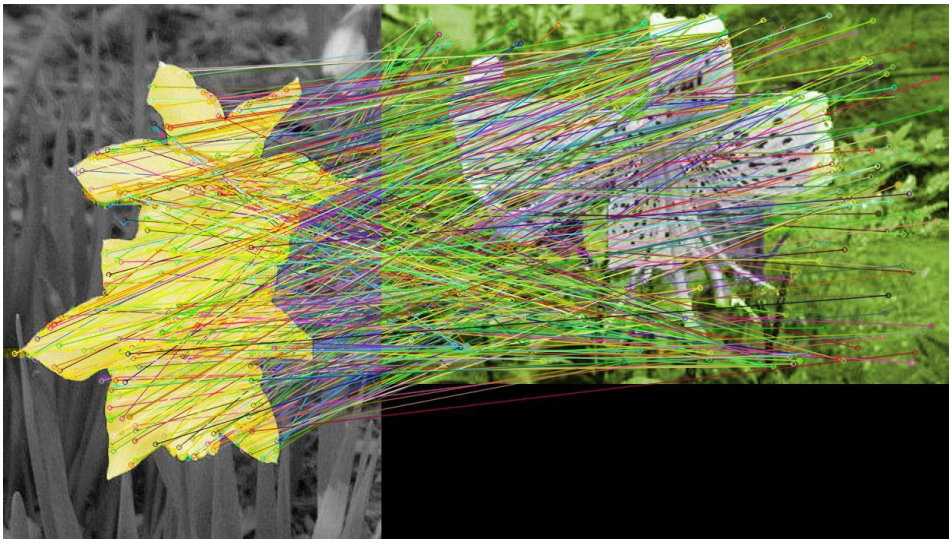


Figura 18: Correspondència entre la regió flor de la imatge a classificar, marcada en groc, i la regió fons de la imatge d'entrenament, marcada en verd.

4.3.3.1. Punts als contorns de les flors

El detector de punts que estem utilitzant per extreure les característiques SURF està generant molts punts d'interès als contorns de la flor, tal com es pot veure a la *Figura 19*. En aquesta figura hem mostrat només els punts que estan prop del contorn de la flor i, com podem veure, alguns dels punts sí que queden a la regió flor (pintada en groc), però molts queden fora, a la regió fons (pintada en morat).

Figura 19: Punts d'interès propers als contorns de la flor.

A més a més, la segmentació realitzada per BPT amb la partició inicial de 20 regions SLIC no defineix amb bona precisió els contorns de les imatges, tal com es pot veure a la mateixa *Figura 19*. Això segurament es podria millorar optimitzant el nombre de regions inicials SLIC, de manera que les regions provinents del BPT s'ajustessin millor als contorns reals de les imatges.

Així doncs, això ens fa correspondre molts punts del contorn de la flor que haurien d'estar a la regió flor i que realment estan col·locats a les regions fons. Per tant, és fàcil que bona part dels punts que estan dintre la flor facin la correspondència amb els punts de contorn que han anat a parar a les regions de fons, fent que aquesta correspondència tingui menor distància i, per tant, sigui triada pel detector.

4.3.4. Ajuntant els punts d'interès per classe i fent farcells de punts

Una vegada fetes les proves tant a nivell d'ajuntar per classe com fent farcells de punts, en aquest punt volíem veure la combinació d'ambdós resultats:

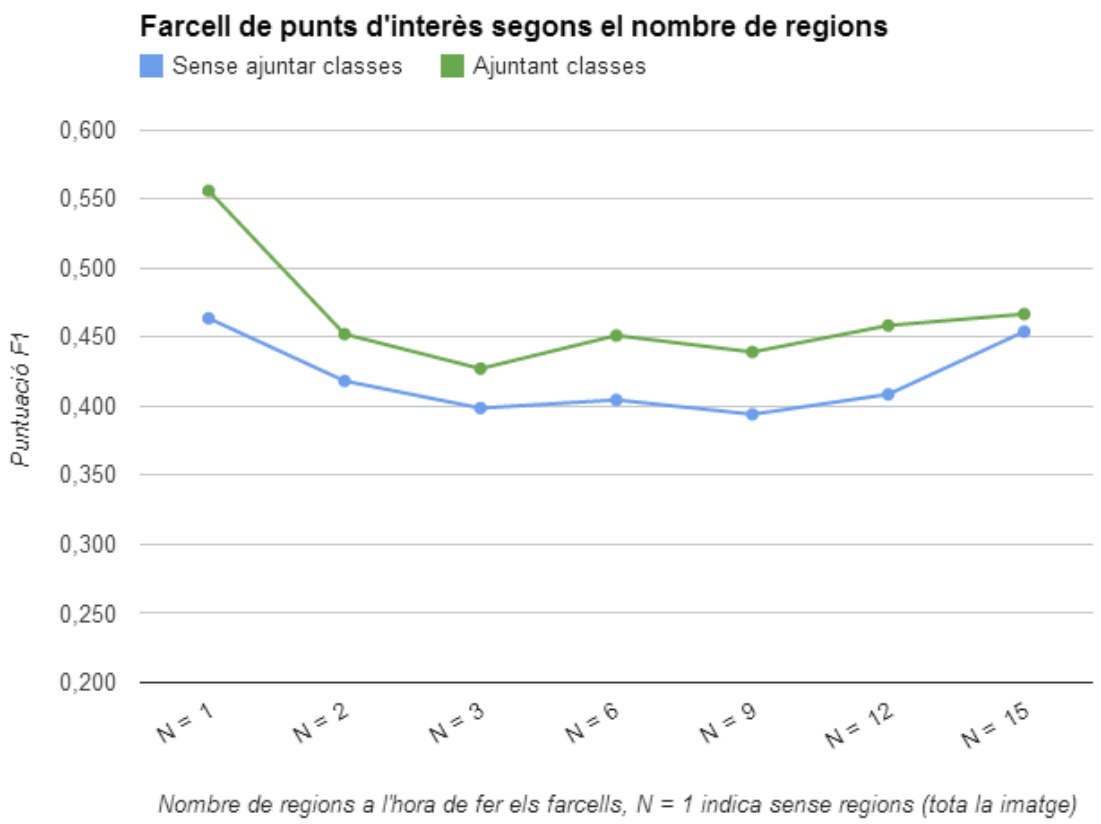


Figura 20: Comparació de la puntuació F1 fent farcells de punts segons el nombre de regions, incorporant el fet d'ajuntar les regions en classes basat en Boiman et al.

Com es pot observar a la *Figura 20* i a la *Taula 4*, ajuntar els punts d'interès en classes millora tots els resultats, però el millor cas és sense fer els farcells de punts (N = 1).

		N = 1	N = 2	N = 3	N = 6	N = 9	N = 12	N = 15
Sense ajuntar en classes	Precision	0,591769	0,539123	0,540683	0,545467	0,531806	0,509213	0,601114
	Recall	0,381372	0,342157	0,315996	0,321569	0,313726	0,341176	0,364706
	F1 score	0,463660	0,418277	0,398642	0,404545	0,394147	0,408593	0,453977
Ajuntant en classes	Precision	0,779526	0,605700	0,597321	0,676157	0,625174	0,617810	0,718368
	Recall	0,432353	0,360784	0,333333	0,339217	0,339216	0,365686	0,347059
	F1 score	0,555967	0,452199	0,427203	0,451188	0,439219	0,458416	0,466692

Taula 4: Resultats de la classificació fent farcells de punts i ajuntant els punts d'interès en classes.

4.3.5. Estudi comparatiu dels resultats obtinguts

Tal com es pot observar a la *Figura 21*, la millor de les configuracions és la que ajuntem els punts d'interès en classes però no fem farcells de punts en regions.

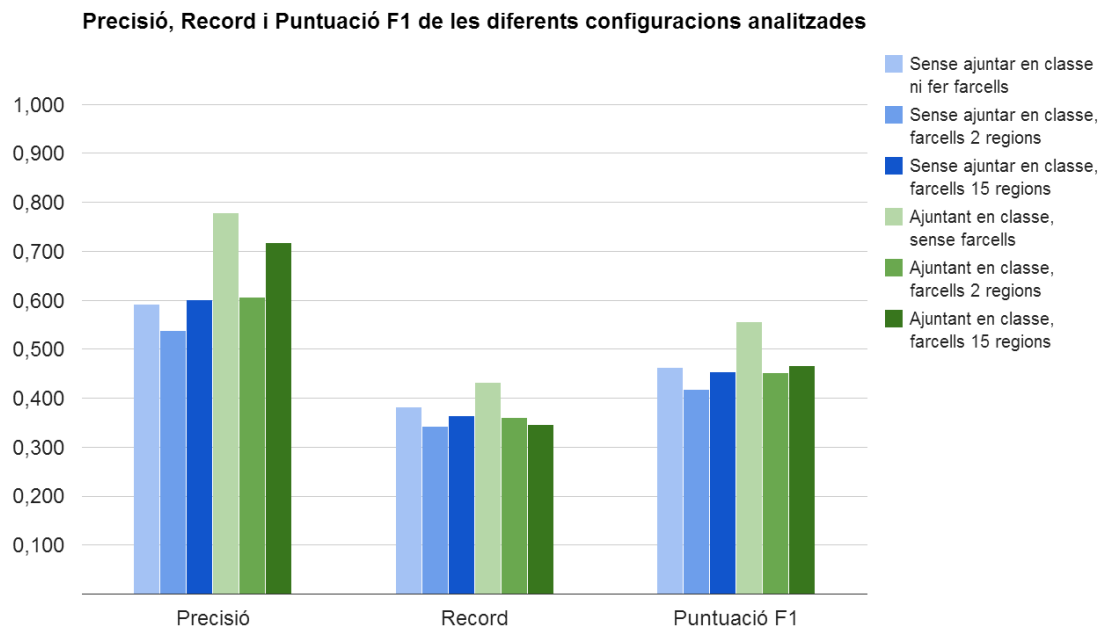


Figura 21: Análisis comparatiu dels resultats segons les diferents configuracions utilitzades.

També podem observar fàcilment com les configuracions en què es fan farcells de dues o quinze regions tenen una pitjor puntuació F1 que els casos en què no fem farcells de punts, cosa que ens indica que aquesta proposta tal com s'ha implementat no millora els resultats.

5. Pressupost

Aquest projecte s'ha desenvolupat amb programari en C++ i Python sota la plataforma ImagePlus del Grup de Processament d'Imatge del Departament de Teoria del Senyal i Comunicacions de la Universitat Politècnica de Catalunya.

En ser un estudi comparatiu, el seu objectiu no és tenir un producte per a l'usuari final, així que no hi ha un cost directe de manteniment.

Per tant, els costos d'aquest projecte vindrien donats per la despesa de personal associada.

	Personal	Preu/hora	Dedicació	Sou	N. mesos	Total
Enginyer junior	1	8,00 €	30h	1.056 €	6	6.336 €
Enginyer senior	2	20,00 €	5h	440 €	6	5.280 €
					TOTAL	11.616 €

Taula 5: Pressupost del projecte.

6. Conclusions i desenvolupament futur

6.1. Conclusions

L'objectiu principal d'aquest projecte era avaluar el potencial d'agrupar punts d'interès d'acord a les regions obtingudes per una segmentació. Com hem pogut veure a l'apartat de resultats, hem complert l'objectiu i hem pogut comprovar com això, com a mínim per una versió senzilla del sistema, no era cert: la puntuació F1 dels sistemes amb diferents agrupaments de punts en farcells de regions és inferior a la configuració on no s'agrupen.

A l'apartat 4. Resultats, pàgina 26, s'han explicat les possibles causes d'aquests valors, que per nosaltres no són intuitius, així que a l'apartat 6.2. Desenvolupament futur s'analitzen els possibles experiments i configuracions que podrien permetre millorar els resultats i veure si aquesta proposta de farcells de punts segons les regions podria ser millor que la configuració bàsica sense farcells.

En no haver obtingut resultats positius, s'ha optat per seguir desenvolupant el programari en comptes de dedicar els esforços a la realització d'un estudi comparatiu amb l'estat de l'art, ja que aquest estudi tindria més sentit després d'un desenvolupament futur.

D'altra banda, s'ha incorporat a l'estudi la idea de Boiman et al [5] d'ajuntar els punts d'interès en classes, de manera que hem pogut comprovar que aquesta configuració sí que millora els resultats respecte la comparació directa imatge contra imatge.

A més a més, com a treball significatiu, s'han desenvolupat diferents eines de programari per al Grup de Processament d'Imatge (GPI) del Departament de Departament de Teoria del Senyal i Comunicacions (TSC). Aquestes eines queden a disposició del GPI per tal de ser utilitzades i millorades. Tal com s'havia especificat, el desenvolupament del programari ha tingut en compte les diferents estructures de comparació (ajuntant o no a nivell de classe i agrupant o no en farcell de punts d'una mateixa regió).

6.2. Desenvolupament futur

6.2.1. Altres configuracions a incorporar a l'anàlisi comparatiu

En aquest anàlisis comparatiu hem estat treballant amb quatre configuracions diferents que tenien en compte dues variables: ajuntar punts d'interès en classes i fer farcells de punts en regions. Es podrien analitzar d'altres variacions de cara a veure com pot millorar la classificació en cadascun dels casos. Com a exemples intuitius:

- Ajuntant en classes diferents imatges a classificar (que sabem que són de la mateixa classe, però no de quina), i ajuntant (o no) en classes les imatges d'entrenament.
- Fent farcells de punts en regions a la imatge a classificar, sense fer farcells de punts a la imatge d'entrenament (aplicant alguna altra restricció espacial).

6.2.2. Altres classificadors i descriptors

En aquest projecte hem volgut fer la proposta a nivell d'ajuntar punts d'interès en classes i d'agrupar-los en farcells de punts segons la seva regió utilitzant els paràmetres més senzills possibles. Es podria millorar el rendiment en desenvolpaments futurs optimitzant els paràmetres (incrementar la k, utilitzar altres descriptors o combinar-ne diversos), així com utilitzant algun altre classificador.

6.2.3. Punts aberrants

A l'hora d'analitzar els resultats, un possible problema que hauríem de comprovar si s'està produint és el dels punts aberrants (en anglès, *outliers*). Quan calculem la distància entre dues regions estem fent un amitjanament entre tots els punts de la regió de la imatge a classificar contra els de la regió de la imatge d'entrenament. Si tenim algun descriptor fora del normal dintre de la regió flor de la imatge a classificar, podem generar una distància massa gran que provoqui que la regió adequada de la imatge d'entrenament sigui descartada.

Una manera de solucionar això podria ser fent la mediana de totes les distàncies, quedar-nos només amb la millor correspondència de cada regió, o bé quedar-nos amb un percentatge de les millors correspondències, per descartar així els punts aberrants.

6.2.4. Deixar d'utilitzar els punts d'interès dels contorns

Una opció que podria ser interessant a l'hora d'evitar que la correspondència entre regions sigui creuada (regió flor que tingui millor correspondència amb regió fons, o viceversa) seria deixar d'utilitzar aquells punts d'interès que siguin massa propers als contorns de les regions.

D'aquesta manera, estaríem assegurant en major mesura que els punts de l'interior de la flor a classificar fessin la correspondència amb els punts de la flor d'entrenament.

6.2.5. Optimitzar la quantitat de regions SLIC a l'hora de generar els BPT

Com s'ha pogut analitzar a l'apartat 4.3.3 de resultats, no totes les segmentacions en regions que hem fet mitjançant els arbres binaris BPT són prou bones. Una prova que es podria fer per tal de millorar els resultats és optimitzar el nombre de regions inicials SLIC a partir del qual es generen aquests arbres, que en el nostre cas era de 20 per tal d'optimitzar el temps dedicat, però es podria augmentar per optimitzar els resultats.

6.2.6. Utilitzar segmentacions perfectes per a les imatges d'entrenament

En aquest sentit, i per millorar encara més les segmentacions, es podrien utilitzar segmentacions perfectes, és a dir, utilitzar les màscares disponibles a la base de dades *17 Category Flower Dataset*. Això ens permetria donar per bona la segmentació utilitzada a les imatges d'entrenament, de manera que segurament tindríem unes correspondències més bones entre regions.

Aquesta opció no l'hem pogut aplicar fàcilment ja que no està disponible la segmentació perfecta per a tota la base de dades, només per a un conjunt de classes, cosa que comportava recalcular els experiments fets per a poder-los comparar i analitzar l'impacte entre utilitzar una segmentació perfecta i una segmentació imperfecta.

6.2.7. Descartar les regions que toquin als marges de la imatge

En moltes ocasions, quan fem una fotografia, acostumem a enquadrar l'element central de manera que no toqui als costats de les imatges. Ho podem veure, per exemple, en els següents casos:

Figura 22: Subconjunt de les imatges en què hi ha un bon enquadrament de les flors.

Així doncs, una possibilitat que podríem tenir en compte és la de descartar les regions que toquen als marges de la imatge, de manera que poguéssim utilitzar gairebé sempre les regions imatge en comptes de les regions fons. Això, tot i així, fallaria en els casos en què les flors estiguessin tocant als contorns.

6.2.8. Utilitzar candidats a objecte

El cas anterior podria ser una manera de detectar els candidats a objecte, però hi ha altres sistemes [15] que permeten trobar candidats a objecte amb resultats bastant bons. Aplicar-los ens hauria de poder permetre millorar el nostre sistema, sempre i quan el fons no estigui aportant informació important a la classificació, cosa que hauríem de comprovar experimentalment.

Bibliografia

[1] M-E. Nilsback & A. Zisserman, «A Visual Vocabulary for Flower Classification» de *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[2] C. Ventura, «Visual Object Analysis Using Regions and Interest Points, PhD Thesis Proposal» de *Departament de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya*.

[3] D. G. Lowe, «Distinctive Image Features from Scale-Invariant Keypoints» de *International Journal of Computer Vision, Volume 60, Issue 2, pp. 91-110*, November 2004.

[4] H. Bay, T. Tuytelaars, L. Van Gool, «SURF: Speeded Up Robust Features» de *9th European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science Volume 3951, pp 404-417*, 2006.

[5] O. Boiman, E. Shechtman, M. Irani, «In Defense of Nearest-Neigbor Based Image Classification» de *CVPR, pp.1-8, 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA. ISBN: 978-1-4244-2242-5*, 23-28 June.

[6] K. López, «Identificació automàtica de càmeres en produccions televisives amb el classificador K-veïns més propers, Projecte Final de Carrera» de *Departament de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya*, 2009.

[7] J. Kim, K. Grauman, «Asymmetric Region-to-Image Matching for Comparing Images with Generic Object Categories» de *In Computer Vision and Pattern Recognition (CVPR). IEEE Conference 2010, pp. 2344-2351*.

[8] Z. Wu, Q. Ke, M. Isard, J. Sun, «Bundling features for large scale partial-duplicate web image search» de *Computer Vision and Pattern Recognition (CVPR). IEEE Conference 2009, pp. 25-32*.

[9] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, Q. Tian, «Building contextual visual vocabulary for large-scale image applications» de *Proceedings of the international conference on Multimedia, ACM, 2010, pp. 501–510*.

[10] OpenCV Dev Team, «OpenCV 3.0.0 documentation, Feature Detection & Description» http://docs.opencv.org/trunk/doc/py_tutorials/py_feature2d/py_matcher/py_matcher.html. [Últim accés: 08 07 2014].

[11] Vilaplana V, Marqués F, Salembier P, «Binary partition trees for object detection» de *IEEE Transactions on Image Processing 2008;17:1–16*.

[12] G. L. Salembier P, «Binary partition tree as an efficient representation for image processing, segmentation and information retrieval» de *IEEE transactions on image processing 2000;9:561–576.*.

[13] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, «SLIC superpixels compared to state-of-the-art superpixel methods» de *Pattern Analysis and Machine Intelligence, 2012 IEEE Transactions on, 34(11), 2274-2282*.

[14] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, «EPFL Technical Report no. 149300» June 2010.

[15] Arbelaez P, Pont-Tuset J, Barron J, Marqués F, Malik J, «Multiscale Combinatorial Grouping» de *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[16] J. Sánchez, «Visual classification with bag of features and nearest neighbours, Bitsearch» http://bitsearch.blogspot.com.es/2013/08/image-recognition-system-classify-mushrooms.html. [Últim accés: 08 07 2014].

Annex

Article "*Visual classification with bag of features and nearest neighbours*" [16] escrit al blog Bitsearch, Research related to the team directed by Xavier Giró-i-Nieto, durant la definició d'aquest Treball Final de Grau. La idea original del projecte era desenvolupar una aplicació de reconeixement de bolets, però finalment vam creure més adequat fer l'experimentació sobre una base de dades ja anotada sobre la qual es podien aplicar les diverses línies d'interès a explorar.

My final grade project is going to be an application which, with an image of a mushroom as the input, is going to classify it and assign a class of mushroom with a certain probability of success.

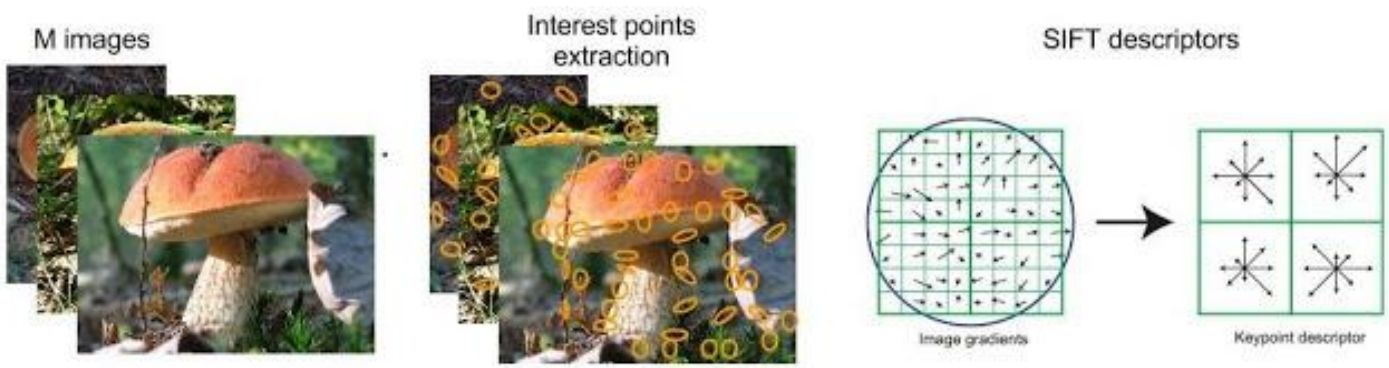
I am going to use supervised learning, so I will need to generate a set of training data with many mushroom images labeled with their mushroom class. The more images labeled, the more precise are going to be the results, so I willl need to determine how many images we need to be as precise as possible but not particularly time-consuming.

- **General scheme of image classification**

My image retrieval algorithm is going to return the results based on a similarity score (distance) between the queryimage signature and the signatures of the database images, ranked accordingly. To generate our signatures of the database images, we have to build a visual vocabulary.

- **Feature extraction**

First of all, we have to build a visual vocabulary for the mushroom images that we are going to identify in our system, so we need to extract features from a set of training images. To extract these features, we are going to consider a set of interest points from each image and after that, apply a SIFT descriptor to those points.



* Mushroom image by Tifred25 (http://commons.wikimedia.org/wiki/File:Bolet_Orange_01.jpg)

To extract the interest points we have two basic methods, dense and sparse sampling:

- Dense sampling: patches of fixed size and shape are placed on a regular grid. It gives a better coverage of the image, a constant amount of features per image area, and simple spatial relations between features. (Tinne Tuytelaars, *Dense interest points*)
- Sparse sampling: focuses on "interesting" locations in the image. Corner detection is an approach used within computer vision systems to extract certain kinds of features and infer the contents of an image. The first corner detector with automatic scale selection is named Harris-Laplace and is one of the most famous techniques. Another one is the SUSAN technique, an acronym standing for *smallest univalue segment assimilating nucleus*, which places a circular mask over the pixel to be tested (the nucleus), and compares every pixel in the mask with the nucleus.

The *Scale Invariant Features Transform* (SIFT) descriptor is one of the most used descriptor when using a bag of features approach. This descriptor generates a 128-dimensional signature of responses to eight oriented gradient filters in a 4x4 grid.

- **Vocabulary construction by feature clustering**

When we have extracted the 128-dimensional SIFT descriptor of every interest point of our image, we will cluster them to generate a vocabulary of k words, as explained at the Video Google paper by Sivic and Zissermann.

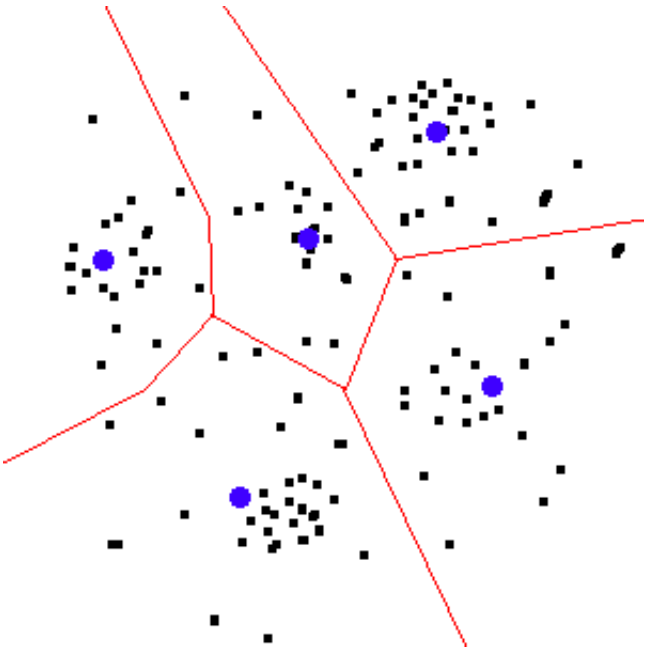
To cluster our data, we will apply the widely adopted K-means algorithm. This clustering method aims at partitioning N observations into K clusters, in which each observation belongs to the cluster with the nearest centroid. The partitions defined by the k-means algorithm are named Voronoi cells. The k-means algorithm is computationally difficult, but there are some efficient algorithms that are usually used and converge quickly to a local optimum.

Given the set of all 128-dimensional SIFT vectors of all of the interest points extracted from our database images $X = \{x_1, x_2, \ldots, x_n\}$, k-means clustering aims to partition the n observations into k sets ($k \leq n$), $S = \{S_1, S_2, \ldots, S_k\}$, minimizing the distance between original vectors and new clustered sets:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

where μ_i is the centroid of the points in S_i. The most common algorithm, called Lloyd's algorithm, uses an iterative technique by alternating between two steps until converging (when the assignments no longer change).

Assignment / Expectation step: Assign each observation to the cluster whose mean is closest to it.



Update / maximization step: Calculate the new means to be the centroids of the observations in the new clusters.

To understand the algorithm, it's a good idea to watch the animation shown at Visuals and Animations website, which is a clustering with K = 4 clusters:

- **Bag of features**

Bag of Features is a model that can be approached by many computer vision tasks, as image classification, video search, robot localization, texture recognition... I am going to implement it in this project to identify and classify photographies of mushrooms taken from a mobile phone.

Taking the definition by Stephen O'Hara and Bruce A. Draper at *Introduction to the Bag of Features Paradigm for Image Classification and Retrieval*:

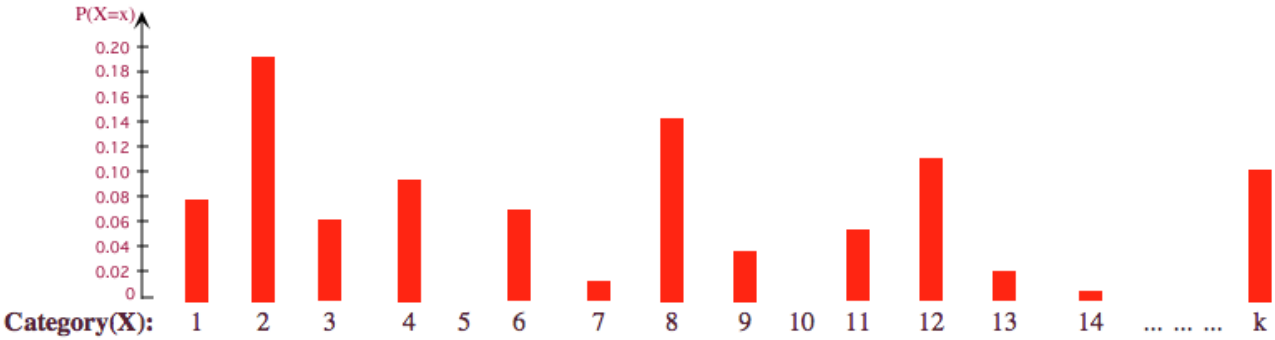
BoF [Bag of Features] methods are based on orderless collections of quantized local image descriptors; they discard spatial information and are therefore conceptually and computationally simpler than many alternative methods.

So, to sum up, it's important to know that bag of features are collections which are orderless and discard spatial information extracting local image descriptors. As we have seen, extracting the SIFT of the points of interest we have lost the spatial information and keep only the image descriptors.

- **Generating the signature**

Once all of our N SIFT descriptors have been clustered into K clusters, it's time to generate a signature of every image in our database.

The signature is a representation of the SIFT descriptors of every interest poing of our single image. We assign every SIFT descriptor to the nearest cluster and we generate the signature by counting how many SIFT descriptors have been assigned to every cluster. If we have K clusters and we normalize our signature dividing by the number of SIFT descriptors, we have a signature like:



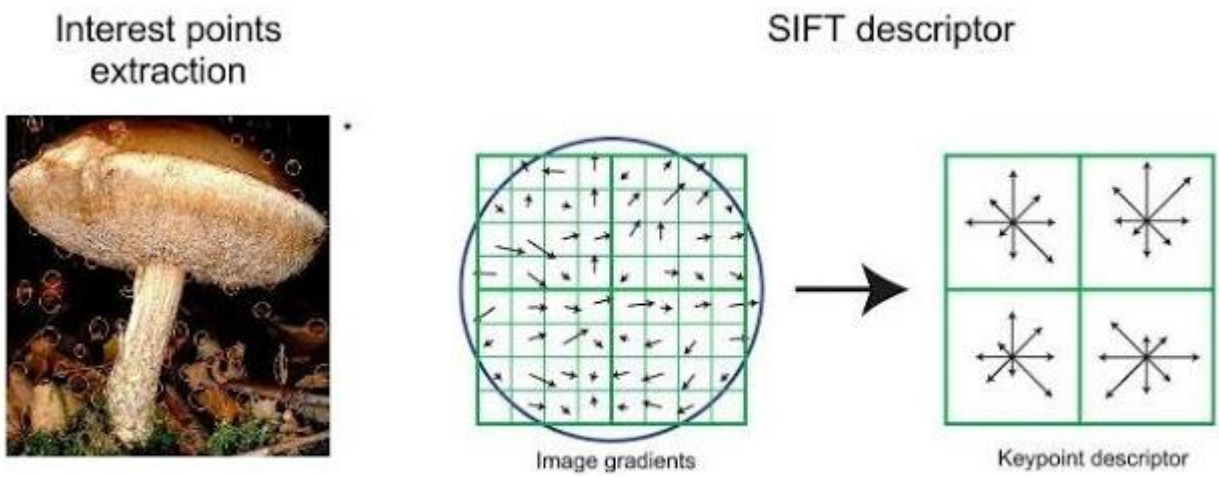
We have to generate a signature for every image in our database in order to compare them with the new signature of a query image in the matching step.

After that, we have to determine the relative importance of every visual word of an image depending on the visual words of other images. That's because some visual words would be more significant than others defining our image. The term frequency - inverse document frequency is a numerical statistic used as a weighting factor which reflects how important a visual word is to a codebook. You can read more about TF - IDF in this previous post by Monica.

- **Classification by nearest neighbour**

Once we have our visual vocabulary (codebook) and the signatures generated by every image, it's time to introduce a new unlabelled image to the system.

As previously done with the training dataset, we have to detect the interest points of this query image and extract the same features (in my case, the 128-dimensional SIFT features for every one of the interest points detected). After that, we have to vector quantize the extracted SIFT features with respect to the generated vocabulary. With this quantization we generate the signature of our image by counting how many SIFT of the input image have been assigned to every cluster.



* Mushroom image by Franc Le Blanc (http://www.flickr.com/photos/olibac/1813130174/)

Once we have the signature of the new input image, it is time to do the matching, by comparing this signature with the signatures of the labelled images in the database. To do that we use Nearest Neighbors, which is going to return a ranked list with the closest signatures in our database according to the distance chosen.

To compare our new query signature to all of our database signatures, we can use different distances, and one of my areas of work will be to determine which distance is the best for my application. I am going to try Euclidean distance , cosine similarity and histogram intersection. I could also try the Manhattan or Mahalanobis, and any other distance explained in the state of the art of image classification with a bag of features approximation.

Vocabulari i glossari

Anotació semàntica: document que assigna a cada imatge d'entrenament de la base de dades la seva corresponent classe semàntica. En anglès, *semantic annotation*.

Classe semàntica: cadascun dels tipus d'objecte que podem tenir en un conjunt. En el nostre cas, cadascun dels tipus de flors. En anglès, *semantic class*.

Correspondència: acció de relacionar un punt d'interès de la imatge a classificar amb el punt més proper d'una imatge d'entrenament. En anglès, *matching*.

Descriptors visuals: característiques detectades sobre dels punts d'interès que representen en certa manera la informació d'aquests punts i els seus voltants. També són anomenats característiques locals. En aquest projecte s'utilitza el descriptor SURF i se cita el descriptor SIFT. En anglès, *visual descriptors* o *local features*.

Farcell de punts: punts que tractem de forma agrupada, en el nostre cas agrupada a nivell de regions. En anglès, *bundle* o *bundling*.

Imatge a classificar: imatge que no està anotada, que és l'entrada del sistema de classificació i de la qual acabem obtenint la classe més probable. També s'anomena imatge de test. En anglès, *query image*.

Imatge d'entrenament: imatge que està anotada a la base de dades amb la seva classe semàntica que ens serveix per entrenar el sistema de classificació. En anglès, *target image*.

Ontologia semàntica: document que relaciona els identificadors de les classes amb el seu nom. En anglès, *semantic ontology*.

Punts d'interès: punts de les imatges que han estat detectats com a interessants per a l'extracció de característiques. En anglès, *interest points*.

Segmentació: distribució d'una imatge en diferents regions. En anglès, *segmentation*.

Superpíxels: regió que conté un conjunt de pixels que ens serveixen per inicialitzar la generació de la jerarquia BPT.

Arbre de partició binària: descomposició d'una imatge en una jerarquia de particions, on per cada nivell de la jerarquia dividim cada node de l'arbre en dues regions. En anglès, *Binary Partition Tree* (BPT).

Veritat Terreny: dades anotades manualment a la nostra base de dades i que donem com a certes per al nostre sistema de classificació. En anglès, *Ground Truth*.