UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

RESUM VISUAL DE SEQÜÈNCIES EGOCÈNTRIQUES MITJANÇANT IMATGES REPRESENTATIVES

Treball de Fi de Grau

Presentat a la Facultat

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

per

Ricard Mestre Subirats

En compliment parcial

dels requisits per al grau en

ENGINYERIA DE SISTEMES AUDIOVISUALS

Dirigit per: Xavier Giró i Nieto

Barcelona, Gener de 2015

Abstract

This Final Degree Work approach the problem of the visual summarization of sets of images captured by an eggocentric camera for lifelogging purposes.

In first place we try to group the images (which represent the day of person's life) into distinguishable and significant esdeveniments. For this purpose, we use visual features extracted with the software Caffe[6].

In second place, we explain the design of extraction techniques of the representative images through similarity graphs.

Finally we analyze the assessment scores given by different users whom we presented the different visual summaries obtained in this project. We achieve a 60% of favorable opinions of the quality of the visual summaries obtained with techniques developed in this project.

Resum

Aquest Treball de Final de Grau aborda el problema de resumir visualment conjunts d'imatges capturats mitjançant una càmera egocèntrica per a propòsits de registre vital (*lifelogging* en anglès).

En primer lloc s'intenta agrupar les imatges (que representen el dia d'una persona) per esdeveniments distingibles i significatius. Per fer-ho s'utilitzen característiques visuals extretes amb el programari Caffe[6].

En segon lloc s'explica el disseny de tècniques de extracció d'imatges representatives mitjançant grafs de similitud.

Per últim s'analitzen les puntuacions d'avaluació donades per diversos usuaris als quals se'ls han presentat els diferents resums visuals obtinguts en aquest projecte. S'ha pogut assolir un 60% d'opinions favorables a la qualitat dels resums obtinguts amb tècniques desenvolupades en aquest treball.

Resumen

Este Trabajo de Final de Grado aborda el problema de resumir visualmente conjuntos de imágenes capturadas mediante una càmera egocéntrica para propositos de registro vital (*lifelogging* en inglés).

En primer lugar se intenta agrupar las imágenes (que representan el día de una persona) por esdevenimentos distinguibles y significativos. Para ello, se utilizan características visuales extraídas con el programario Caffe[6].

En segundo lugar se explica el diseño de técnicas de extracción de imágenes representativas mediante grafos de similitud.

Por último se analizan las puntuaciones de avaluación dadas por diversos usuarios a los que se les ha presentado los diferentes resúmenes visuales obtenidos en este proyecto. Se ha conseguido alcanzar un 60% de opiniones favorables a la calidad de resúmenes obtenidos con técnicas desarrolladas en este trabajo.

Agraïments

Voldria agrair primer de tot al grup que m'ha donat suport per poder elaborar aquest treball, el qual sense ells i elles no hagués estat posible: Xavi Giró, Petia Radeva, Estefanía Talavera i Marc Bolaños.

En segon lloc voldria agrair al grup d'imatge de l'UPC l'oportunitat de treballar amb un programari molt potent, i als seus técnics per l'ajuda prestada quan certes parts no funcionaven.

També voldria agrair a totes les persones que van contestar els qüestionaris d'avaluació. Sense ells no hauriem tingut cap resultat.

Per acabar, voldria agrair a la meva familia i els meus amics (vells i nous) el suport que m'han donat durant tot el grau. Sense la vostra inestimable companyia no m'hagués estat possible!

Moltissímes gràcies a tots!

Historial i aprovació de revisions

Revisió	Data	Propòsit
0	18/01/2015	Creació del document
1	26/01/2015	Revisió del document per part d'en Xavier Giró
2	03/02/2015	Revisió del document per part d'en Ricard Mestre
3	04/02/2015	Revisió del document per part d'en Xavier Giró

LLISTA DE DISTRIBUCIÓ DEL DOCUMENT

Nom	e-mail
Ricard Mestre Subirats	rcrd.mstr@gmail.com
Xavier Giró i Nieto	xavier.giro@upc.edu

Escrit per:		Revisat i aprovat per:	
Data	18/01/2015	Data	04/02/2015
Nom	Ricard Mestre Subirats	Nom	Xavier Giró i Nieto
Posició	Autor del projecte	Posició	Supervisor del projecte

Taula de continguts

Abstract 1
Resum 2
Resumen 3
Agraïments 4
Historial i aprovació de revisions 5
Taula de continguts 6
Llista de figures 9
Llista de taules 10
1. Introducció i objectius 11
2. Requeriments i especificacions 12
2.1. Col·laboració amb la Universitat de Barcelona 12
2.2. Requeriments tècnics 12
2.3. Requeriments d'entorn 12
2.3.1. Base de dades 13
2.4. Origen del projecte 13
2.5. Pla de treball 13
2.5.1. Work packages 13
2.5.2. Incidències i desviacions 13
2.5.3. Diagrama de Gantt 14
3. Estat de l'art 17
3.1. Efficient Retrieval from Large-Scale Egocentric Visual Data Using a Sparse Graph Representation 17
3.2. Video segmentation of life-logging videos 17
3.3. A Survey on Recent Advances of Computer Vision Algorithms for Egocentric Video 18
3.4. Story-Driven Summarization for Egocentric Video 18
4. Metodologia: 19
4.1. Vista general de la solució: 19
4.2. Extracció de característiques 19
4.3. Normalització 20
4.4. Clusterització d'imatges per similitud visual 20
4.4.1. Average 21
4.4.2. Centroid 21

4.4.3. Complete 22
4.4.4. Median 22
4.4.5. Single 23
4.4.6. Ward 23
4.4.7. Weighted 23
4.5. Divisió de clústers visuals per consistència temporal 24
4.6. Fusió de mini-clústers amb els seus veïns temporals 24
4.7. Extracció de l'imatge representativa 25
4.7.1. Passejada aleatòria 26
4.7.2. Distància total mínima 26
5. Resultats 28
5.1. Avaluació de la clusterització 28
5.1.1. Índex de Jaccard 28
5.1.2. Efecte dels métodes d'enllaç 29
5.1.3. Comportament general 29
5.1.4. Efecte de la normalització 30
5.1.5. Efecte de la divisió-fusió 31
5.2. Avaluació de l'extracció d'imatge representativa 33
5.2.1. Avaluació de la representativitat individual 34
5.2.2. Avaluació de la representativitat conjunta 38
6. Pressupost 42
7. Conclusions i treball futur: 43
7.1. Conclusions 43
7.2. Treball futur 43
8. Bibliografía: 45
9. Annexos: Developer guide 47
9.1. Visual summary extraction functions 47
9.1.1. mainVisualSummary 47
9.1.2. extractCNNFeatures 47
9.1.3. extractNF 47
9.1.4. normalizeL2 48
9.1.5. obtainAutomaticCluster 48
9.1.6. orderClustersTemporally 48
9.1.7. orderClusters 49

9.1.8. randomWalkClustering 49
9.1.9. minimumDistanceClustering 50
9.2. Evaluation functions 50
9.2.1. resultsClusterValidation 50
9.2.2. analizarExcel_Narrative 51
9.2.3. JaccardIndex 51
9.3. Other functions 51
9.3.1. resizeImages 51

Llista de figures

Figura 1. Imatge d'exemple d'una càmera Narrative.....11
Figura 2. Primera part del diagrama de Gantt.....15
Figura 3. Segona part del diagrama de Gantt.....16
Figura 4. Diagrama de funcionament general del projecte.....19
Figura 5. Exemple de dos grups.....20
Figura 6. Càlcul de l'enllaç *average*.....21
Figura 7. Càlcul de l'enllaç *centroid*.....22
Figura 8. Càlcul de l'enllaç *complete*.....22
Figura 9. Càlcul de l'enllaç *single*.....23
Figura 10. Exemple d'event agrupat.....25
Figura 11. Exemple d'imatges representatives. A l'esquerra *random walk*, al centre distància total minima i a la dreta imatge aletòria.....26
Figura 12. Exemple de resum visual del dia.....27
Figura 13. Índex de Jaccard, comparació dels mètodes d'enllaç.....29
Figura 14. Índex de Jaccard general.....30
Figura 15. Índex de Jaccard, característiques normalitzades i no normalitzades.....31
Figura 16. Índex de Jaccard, ordenat i sense ordenar.....32
Figura 17. Índex de Jaccard, ordenat (no temporalment) i ordenat temporalment.....33
Figura 18. Exemple de presentació d'un event.....34
Figura 19. Exemple de presentació de les imatges representatives de l'event.....35
Figura 20. Validacions de les tècniques, individualment.....36
Figura 21. Exemple d'un event automàtic amb imatges de dos events.....37
Figura 22. Tria de tècniques com a representativa.....38
Figura 23. Exemple de presentació d'un resum visual complet.....39
Figura 24. Pregunta final.....39
Figura 25. Validacions de les tècniques com a resums visuals.....40
Figura 26. Tria del millor resum visual.....41

Llista de taules

Taula 1. Cost del personal associat.....42
Taula 2. Cost de la llicència del Matlab.....42
Taula 3. Cost total.....42

1. Introducció i objectius

En els darrers anys han sorgit diversos dispositius, com les Google Glasses, la càmera Narrative (figura 1) o la càmera SenseCam de Microsoft, capaços d'enregistrar el dia d'una persona mitjançant una seqüència d'imatges. Aquesta aplicació s'anomena registre vital (en anglès, "lifelogging") i consisteix en enregistrar moments del dia d'una persona desde un punt de vista egocèntric, és a dir, les imatges pretenen simular on mira la persona en cada moment per a després compartir-les a les xarxes socials o simplement per guardar un record de cada dia.

Figura 1. Imatge d'exemple d'una càmera Narrative.

Aquests dispositius son capaços de capturar unes 2000-2500 imatges per dia. De totes aquestes imatges, n'hi pot haver de desenfocades o de irrellevants considerant que hi ha un alt índex de redundància. Imaginem que quedem amb un amic per prendre un cafè, és ben possible que hi estem una hora. En aquest temps el nostre dispositiu pot haver capturat unes 30 o 40 imatges practicament iguals. Si el que volem és guardar un record de la trobada, amb una imatge que representi aquest esdeveniment seria suficient, i per tant la resta d'imatges són redundants.

També s'està utilitzant el registre vital com ajuda per a persones amb problemes de memòria, com els malalts d'Alzheimer. S'està estudiant si ensenyar un resum del dia al malalt en qüestió pot comportar una millor retenció dels records.

Per totes aquestes raons el primer objectiu d'aquest Treball de Fi de Grau és la selecció automàtica de les imatges representatives del dia d'una persona, basant-se en la similitud visual de les imatges.

Durant el dia d'una persona ocorren un cert nombre d'esdeveniments com: el viatge amb autobus desde casa al lloc de treball, el matí treballant, l'aturada per esmorzar... El cas ideal es desaria guardar una imatge de cadascun d'aquests esdeveniments de cada dia. Per aquesta raó el segon objectiu d'aquest treball és el·laborar un algorisme de clusterització d'imatges per esdeveniments i representar cadascun d'aquests amb una imatge seleccionada també automàticament.

2. Requeriments i especificacions

2.1. Col·laboració amb la Universitat de Barcelona

En aquest treball no només hi hem actuat el professor Giró i jo mateix des del Grup de Processament de la Imatge (GPI) de la Universitat Politècnica de Cataluna (UPC), sino que hi hagut una col·laboració amb el grup BCNPCL (Barcelona Percepture Computer Laboratory) de la Universitat de Barcelona, mitjançant la professora Petia Radeva i els seus estudiants Estefanía Talavera i Marc Bolaños.

La col·laboració ha consistit en tres parts: reunions setmanals per a fer el seguiment del projecte i proposar idees i alternatives; l'ús de les imatges obtingudes per la càmera Narrative que pertany al departament; i finalment, l'ús d'alguns fragments de codi que m'han facil·litat l'implementació del treball.

2.2. Requeriments tècnics

L'objectiu de l'algorisme de clusterització que s'ha implementat és separar per esdeveniments totes les imatges del dia, basant-se amb la similitud visual de les mateixes. Dins d'aquesta tasca hi han dos parts ben diferenciades, la clusterització automàtic de les imatges tenint en compte les característiques visuals, i un algorisme de divisió-fusió que intenta corregir errades de la clusterització visual tenint en compte la informació temporal. Per tant tenim dos requeriments:

- Per una banda, obtenir grups d'imatges molt semblants visualment.
- Per l'altra, tenir en compte l'informació temporal per millorar el resultat de l'agrupament.

Pel que fa a l'extracció de l'imatge representativa, en aquest treball ens hem decantat per considerar la imatge més repetitiva visualment parlant, com la més representativa d'un esdeveniment. En l'apartat de treball futur al final del document, discutirem breument sobre altres possibles criteris.

2.3. Requeriments d'entorn

Tot el projecte ha estat implementat amb codi Matlab per dues raons: primer per la meva familiaritat envers de la nul·la experiència amb llenguatge Python, que era l'alternativa. I segon, pel fet de que el grup BCNPCL de l'UB tenia codi ja fet amb Matlab i això podria facilitar certes tasques com l'avaluació de l'agrupament.

Pel que fa a la presa d'imatges per a dur a terme els experiments, s'ha fet amb la càmera Narrative que poseeix el departament de la professora Radeva. Aquesta càmera captura una imatge cada minut aproximadament, tot i que també captura imatges si se li donen dos petits tocs amb el dit seguits, o si nota un canvi de moviment molt sobtat.

2.3.1. Base de dades

La imatges que hem fet servir durant aquest treball són cinc conjunts diferents (de 1388, 721, 586, 684 i 625 imatges), corresponents a cinc dies de captura de la càmera Narrative del grup BCNPCL (Barcelona Percepture Computer Laboratory) de UB.

2.4. Origen del projecte

Aquest projecte té l'origen en un contacte entre el profesor Xavier Giró de la UPC (Universitat Politècnica de Catalunya) i la professora Pètia Radeva del grup BCNPCL de la UB (Universitat de Barcelona). L'actual projecte de la professora Radeva és sobre registre vital i el profesor Giró ho va trobar prou interessant com per proposar-me fer una col·laboració. Degut a que el professor Giró havia treballat abans amb altres projectes sobre l'extracció d'imatges clau per resumir resultats de cerca (PFC Mónica Alfaro[2]) o de videos (PFC David Almendros[1], PFC Manel Martos[9], article al CBMI 2013 [16]) al final ens vam posar d'acord els tres sobre el tema i l'objectiu del meu TFG: selecció imatges representatives, del dia d'una persona.

La professora Radeva compta amb una càmera Narrative per a fer els seus experiments. La Narrative és una petita càmera, d'uns pocs centímetres de diàmetre, que es porta posada o bé penjada del coll, o bé enganxada al coll de la camisa.

2.5. Pla de treball

2.5.1. Work packages

WP1: Proposta de projecte i proposta de treball

WP2: Lectura de l'estat de l'art

WP3: Generació d'una col·lecció anotada d'imatges preses des de Narrative

WP4: Desenvolupament del programari

WP5: Revisió crítica del projecte

WP6: Millora del programari

WP7: Avaluació dels resultats

WP8: Redacció i presentació del Treball de Fi de Grau

2.5.2. Incidències i desviacions

Al principi del projecte, es va contemplar la possibilitat d'utilitzar tècniques de detecció de candidats a objecte com a eina principal per a l'elaboració dels resums visuals, però

degut a les incidències que descriurem a continuació, aquesta part del projecte ha quedat com a treball futur, ja que no ha donat temps de realitzar-la.

Com ja es va comentar a la revisió crítica, vam tenir problemes amb l'instal·lació del programari Caffe, en concret la seva adaptació a Matlab, als equips informàtics del grup d'imatge de l'UPC. Un cop instal·lat el programari i funcionant, vam tenir un procés d'aprenentatge que va durar més del que estava previst, i com a conseqüència es van anar endarrerint algunes tasques.

Més tard, quan encara contemplavem la detecció de candidats a objecte com una part del projecte, vam voler instal·lar uns descriptors de regions de l'imatge, desenvolupats per els mateixos autors de Caffe. Amb aquesta instal·lació també vam tenir problemes, el que va fer, que de cara a la revisió crítica, deixéssim com a opcional la tasca d'aplicar candidats a objecte i només com a eina de millora de resultats.

Veient que probablement no tindriem temps d'aprendre a utilitzar aquests candidats a objecte, vam desestimar definitivament aquesta part del projecte per centrar-nos en millorar els resultats mitjançant l'algorisme de divisió-fusió que descrivim a la secció de desenvolupament del projecte.

També cal comentar que el work package 4 es va eliminar ja que, les tasques d'anotació d'imatges han recaigut en el grup BCNPCL de l'UB.

2.5.3. Diagrama de Gantt

Les Figures 2 i 3 mostren el diagrama de Gantt del projecte.

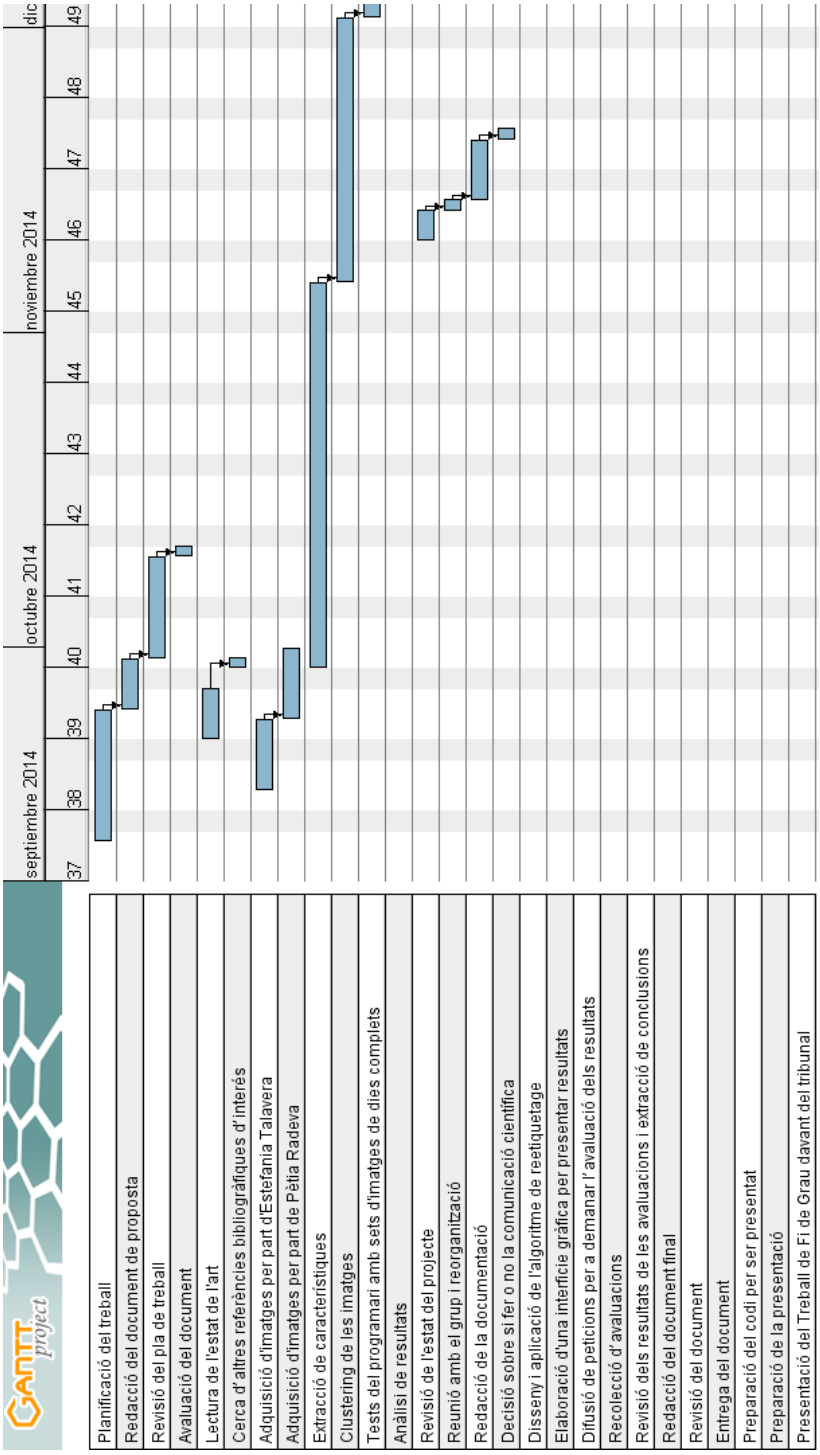


Figura 2. Primera part del diagrama de Gantt.

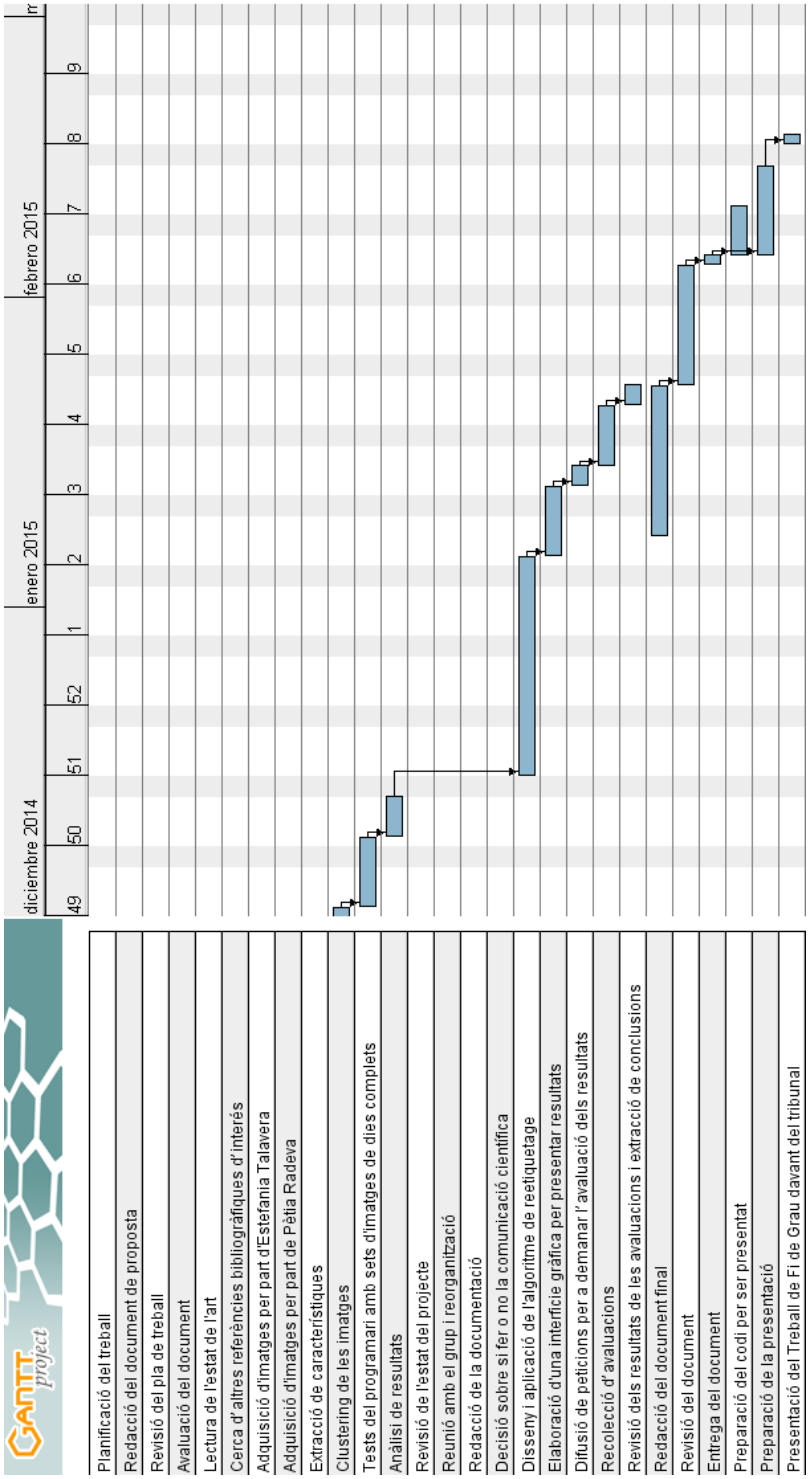


Figura 3. Segona part del diagrama de Gantt.

3. Estat de l'art

En aquest apartat descriurem i compararem propostes d'altres autors que tracten la mateixa temàtiques d'aquest treball.

L'objectiu principal del treball és obtenir un resum visual del dia. Una possible solució a aquest problema és utilitzar grafs de similitud. Un graf de similitud aplicat a el processat d'imatge és un conjunt d'imatges conectades entre si per arestes, on aquestes arestes tenen diferents pesos depenent de la similitud visual que hi hagi entre les imatges que actuen de nodes.

3.1. Efficient Retrieval from Large-Scale Egocentric Visual Data Using a Sparse Graph Representation

La primera proposta basada en grafs presenta una solució per a reduir la gran quantitat d'imatges que s'obtenen a la captura de vídeos de registre vital.

L'autor del paper, Vijay Chandrasekhar[3], defensa que la mida de les dades capturades am vídeos de lifelogging és massa gran (parla de deu hores de vídeo a trenta imatges per segon), i que és pot aprofitar l'alta redundància de les imatges capturades per reduir el volum de dades.

Per fer-ho apliquen grafs de similitud on les arestes valen 1 si hi ha algun tipus de transformació geomètrica entre els nodes. Els autors també van provar de donar pesos entre 0 i 1 però els resultats finals no milloraven tant com per a plantejar-se augmentar el cost computacional.

Per extreure el pes de cada aresta s'apliquen punts d'interés Diference-of-Gaussian (DoG) i descriptors visuals SIFT. Després aplicaven la técnica de Bag-of-Words per trobar coincidencies. Un cop obtinguts els grafs s'aplica un algorisme de clusterització i s'extreuen els nodes més representatius, fet que redueix la redundància de la base de dades.

3.2. Video segmentation of life-logging videos

Aquesta proposta encara el problema de la segmentació d'esdeveniments de seqüències de registre vital. Els autors, entre ells els membres del grup BCNLCP amb el qual em col·laborat Marc Bolaños i Petia Radeva[14], es basen en el moviment de la persona que porta la càmera per a segmentar la seqüència en esdeveniments significatius.

Classifiquen les imatges en tres grups: *estàtiques* (la persona i la càmera estàtics), en *trànsit* (la persona movent-se o corrent) i *movent la càmera*(si la persona està estàtica però la càmera no). Al ser una seqüència de baixa resolució temporal, les tècniques d'analisi de video no serveixen i per tant es decanten per combinació de característiques com borrositat i color, i aplicant tècniques de grafs de similitud per a classificar aquestes característiques.

3.3. A Survey on Recent Advances of Computer Vision Algorithms for Egocentric Video

Aquest article ens proporciona moltes referències a diferents treballs sobre captura de video egocèntrica. En concret per al registre vital, fa un resum sobre les propostes en resum de vídeo, detecció de novetats i interaccions socials.

En primer lloc, en resum de vídeo ens explica que la proposta de Doherty[10] fa una segmentació d'esdeveniments, igual que hem dissenyat en aquest treball, i després també es basa en distàncies entre les imatges per obtenir les imatges representatives de cada esdeveniment. Alguna tècnica que utilitzen es agafar la imatge mitjana del esdeveniment, o una tècnica més elaborada és la d'agafar l'imatge més propera a la mitjana de tots els esdeveniments. En canvi la proposta de Lee [11] es basa en la idea que una imatge amb contingut probablement serà més representativa que la que es basa en distància visual. Per obtenir aquesta imatge busquen cares, mans o objectes.

En segon lloc, per a la detecció de novetats s'analitza la proposta de Aghazadeh[12]. L'autor defensa que una forma de detectar novetats al dia d'una persona és pot obtenir comparant les imatges de molts de dies, i qualsevol esdeveniment que alteri la rutina implicarà una novetat candidata a ser recordada.

Pel que fa a les interaccions socials, Fathi[13] es basa en la detecció de cares per detectar esdeveniments socialment memorables de l'usuari de la càmera. Utilitza diversos mètodes per calcular l'atenció que li estan prestant les persones que miren a l'usuari i depenent d'aquesta atenció classifiquen l'interacció com un diàleg, una discussió, un monòleg o d'altres categories.

3.4. Story-Driven Summarization for Egocentric Video

L'última aproximació, i probablement la que s'assembla més a aquest treball, és la proposta per a resumir videos de lifelogging de Zheng Lu i Kristen Grauman[4]. Els autors busquen una solució diferent a les plantejades anteriorment que es basaven en la diversitat dels keyframeso en la inclusió de anomalies.

El primer pas de la solució proposada és segmentar el vídeo en esdeveniments basant-se en si la persona que porta la càmera està estàtic, en moviment o movent el cap. Per a aconseguir-ho extreuen característiques de flux òptic i també de borrositat (blurriness en anglès). Un cop trobats els esdeveniments, s'aplica un detector d'objectes per a la següent etapa, que és la de valorar si un esdeveniment explica una història interessant. Per discriminar això es mira si els objectes detectats a una imatge influeixen a la següent, intentant trobar cadenes més grans.

Finalment la proposta optimitza una funció d'energia que dona una puntuació a cada cadena candidata a entrar al resum del dia d'acord a la influència que té al llarg del vídeo i els esdeveniments individuals importants que conté.

4. Metodologia:

En aquesta secció descriurem l'estructura de la solució implementada i es discutiran diferents alternatives disponibles a cada bloc de l'estructura.

4.1. Vista general de la solució:

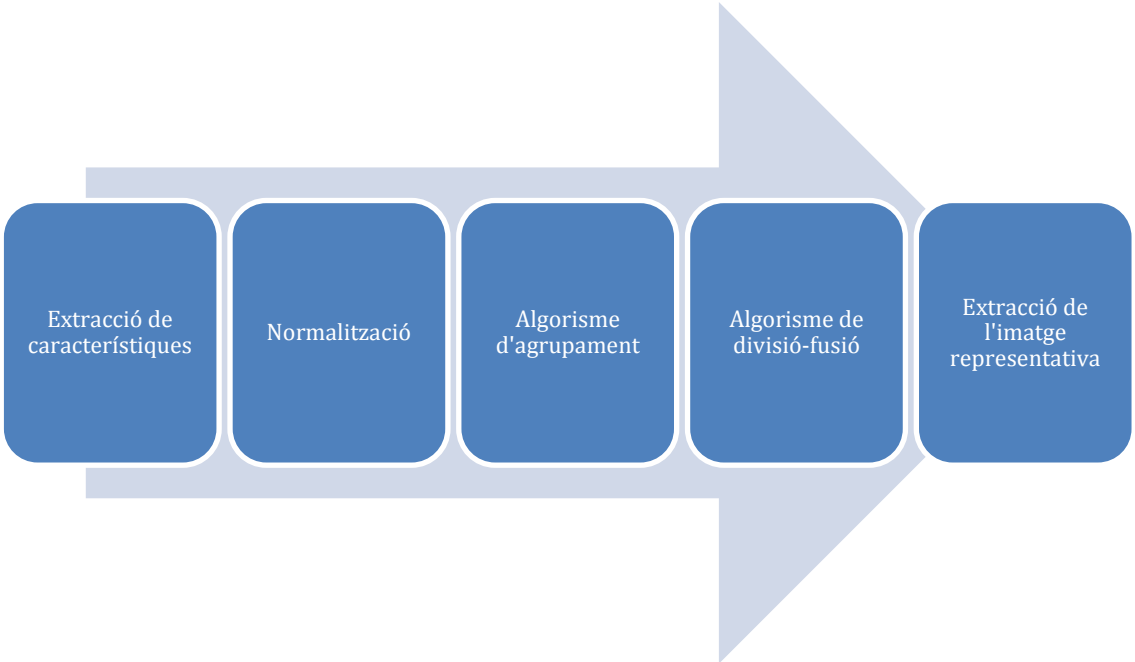


Figura 4. Diagrama de funcionament general del projecte.

4.2. Extracció de característiques

La primera etapa del disseny és l'extracció de característiques. Per poder analitzar totes les imatges necessitem un descriptor visual de les mateixes. En aquest treball s'ha optat per extreure característiques CNN (Convolutional Neural Networks) mitjançant el programari Caffe[6], que ja ve entrenat amb base de dades Imagenet[7].

El Caffe utilitza la técnica d'aprenentatge profund ("deep learning" en anglès) derivat de l'aprenentatge de màquines més general. Dins de moltes formes d'aplicar l'aprenentatge profund en aquest treball s'han aplicat les Xarxes Neuronals Convolucionals (CNN), ja que són invariants a la translació.

Tot i no ser invariants a la rotació les hem pogut utilitzar igual gràcies a que els companys del grup BCNPCL de la UB van desenvolupar un algorisme per orientar les imatges rotades respecte al terra.

El programari ens proporciona un vector de 4096 característiques visuals per a cada imatge que utilitzarem per agrupar les imatges per esdeveniments i per extreure la més representativa.

4.3. Normalització

Un cop extretes les característiques apliquem una normalització a cada vector de característiques per la seva norma:

$$\hat{x} = \sqrt{\sum_{i=1}^{n} |x_k|^2}$$

Aquesta operació comporta una millora en els resultats, com es pot veure a l'apartat de l'avaluació. Aquesta etapa serveix per posar en valors comparables tots els vectors de característiques i així poder classificar més eficientment.

4.4. Clusterització d'imatges per similitud visual

L'objectiu principal d'aquesta etapa és aconseguir agrupar les imatges de tot el dia per esdeveniments. Com que no coneixem el nombre d'esdeveniments del nostre conjunt d'imatges, necessitem un algorisme de clusterització no supervisada. Existeixen moltes alternatives però en aquest treball ens hem decantat per aplicar tècniques de clusterització aglomeratiu, ja que amb la funció *clusterdata* nativa de Matlab hem pogut fer moltes proves fàcilment.

El funcionament general de l'agrupament algomeratiu és el següent: al principi del procés, cada element és un grup en si mateix. Els grups (o clústers) son combinats seqüencialment en clústers més grans fins que tots els elements acaben formant part del mateix clúster. A cada etapa de l'agrupament, els dos clústers separats per la distància més curta s'ajunten. La definició de "distància més curta" és que diferencia entre si els diferents mètodes d'enllaç de l'agrupament aglomeratiu.

Imaginem que tenim dos grups com els següents:

Figura 5. Exemple de dos grups.

Tenim diversos mètodes d'enllaç, és a dir, diversos mètodes per mesurar la distància, sempre euclidiana, entre aquests dos grups i que descriurem en les següents seccions.

4.4.1. Average

Es calculen les distàncies euclidianes de tots els punts de cada clúster amb tots els punts de l'altre i es fa la mitjana de totes aquestes distàncies:

$$distance(A,B) = \frac{1}{|A| \cdot |B|} \sum_{i=0}^{n_A} \sum_{j=0}^{n_B} d(x_i, y_j)$$

On d(x,y) és la distància euclidiana entre els punts.

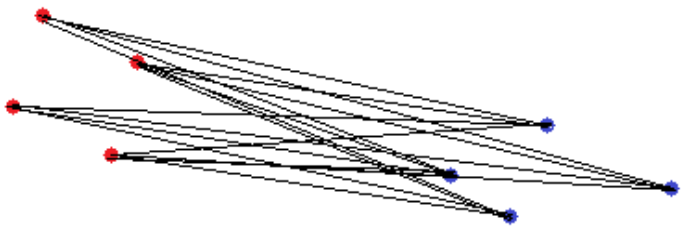


Figura 6. Càlcul de l'enllaç *average*.

4.4.2. Centroid

Es calculen els centroides de cada clúster i es mira la distància entre ells:

$$distance(A,B) = \|\bar{x}_A - \bar{x}_B\|_2$$

On:

$$\bar{x}_A = \frac{1}{n_A} \sum_{i=0}^{n_A} x_{Ai}$$

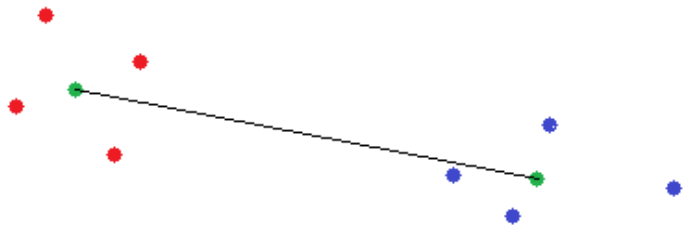


Figura 7. Càlcul de l'enllaç *centroid*.

4.4.3. Complete

S'agafa com a distància la major entre els dos punts més allunyats entre si dels dos clústers.

$$distance(A, B) = max\{d(x, y): x \in A, y \in B\}$$

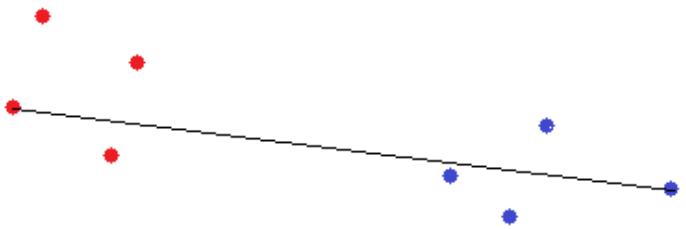


Figura 8. Càlcul de l'enllaç *complete*.

4.4.4. Median

És com el "centroid" però calculant el centroide amb pesos:

$$distance(A, B) = \|\bar{x}_A - \bar{x}_B\|_2$$

On, si considerem que el clúster A s'ha format al ajuntar els clústers C i D:

$$\bar{x}_A = \frac{1}{2}(\bar{x}_C + \bar{x}_D)$$

4.4.5. Single

És l'enllaç complementari al "complete", per tant ara s'agafa la mínima distància entre clústers:

$$distance(A, B) = min\{d(x, y): x \in A, y \in B\}$$

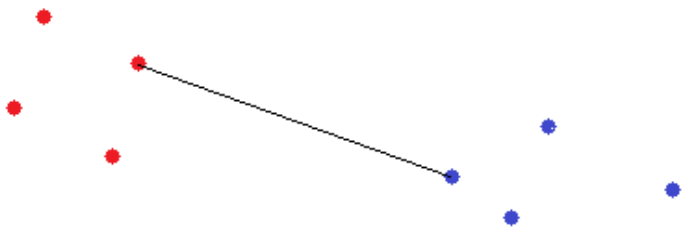


Figura 9. Càlcul de l'enllaç *single*.

4.4.6. Ward

És un mètode que busca que el clúster resultant de la unió sigui el que menys augmenti la variança entre tots els punts del nou clúster. A la primera iteració s'agafa com a criteri la distància euclidiana.

4.4.7. Weighted

Fa servir distàncies mitjanes com *average* però afegint-hi pesos. Per exemple, tenim un clúster A que prové d'ajuntar els clústers C i D. La distància d'aquest clúster amb el clúster B és la següent:

$$distance(A, B) = \frac{1}{2}(distance(B, C) + distance(B, D))$$

A l'apartat d'avaluació (secció 5.1) s'analitza com s'ha escollit el millor mètode d'enllaç, que ha resultat ser l'*average*.

Com ja em dit abans aquest algorisme va ajuntant grups fins que tots acaben en un de sol, però nosaltres, obviament, no volem això. Per tant cal indicar-li on parar de ajuntar. Aquest fet es controla amb el paràmetre *cutoff*, que si val entre 0 i 2, la funció calcularà un nombre de clústers automàtic, i si val més de 2, la funció ho enten com el nombre de clústers que volem. En aquest treball ens decantem per l'opció no supervisada, és a dir, la primera. En l'apartat d'avaluació veurem quin és el valor òptim d'aquest paràmetre.

4.5. Divisió de clústers visuals per consistència temporal

La similitud visual no és un criteri suficient en molts casos per crear esdeveniments per un diari personal. En moltes situacions esdeveniments llargs queden interromputs temporalment per un altre esdeveniment, típicament, més petit. Per exemple, un dia de jornada laboral en un despatx té probablement diversos petits talls quan el treballador fa una pausa per menjar o estirar les cames.

La *càmera Narrative* dona informació temporal de quan ha estat presa la imatge concreta, assignant a cada imatge amb la marca temporal (en anglès, *timestamp*) de quan ha estat capturada. Amb aquesta dada és possible avaluar la coherència temporal dels esdeveniments del dia. Així doncs, si un clúster visual presenta una seqüència de marques temporals que està interrompuda per les marques temporals d'un altre esdeveniment, el primer esdeveniment és dividit en diferents esdeveniments amb segments temporals consecutives.

Aquest és el mateix criteri que s'ha utilitzat també durant l'anotació manual per generar la veritat terreny. En aquesta veritat terreny, es consideren diferents esdeveniments visualment similars però que estan separats per petits esdeveniments de 6 imatges o més.

4.6. Fusió de mini-clústers amb els seus veïns temporals

Si tenim en ment la resolució temporal de la seqüència d'imatges, és a dir, el temps entre imatges capturades amb la càmera *Narrative* (que és d'un minut de mitjana), podem interpretar que un grup de menys de 5 imatges, equival a un esdeveniment de menys de 5 minuts, com anar al lavabo per exemple. En aquest treball em considerat que esdeveniments de menys de 5 minuts (i per tant de menys de 5 imatges), no són representatius del dia, i per tant no els considerem un esdeveniment en si.

En l'etapa anterior de clusterització, la mida dels grups d'imatges resultants varia molt, i sovint n'hi ha de 5 o menys imatges. Per asegurar-nos que aquest fet no afecta al nostre resum visual, hem dissenyat un algorisme de fusió que s'aplica només a aquest clústers que no arriben a una mida mínima. L'objectiu principal d'aquest algorisme és reetiquetar imatges o grups d'imatges menors de 5 i assignar-les a esdeveniments més grans, és a dir, fusionar els esdeveniments petits amb els seus veïns més grans.

El criteri per tal d'assignar els mini-clústers a un dels seus dos veïns més propers s'ha basat únicament en criteris temporals, assignant-lo al clúster més proper. La proximitat temporal entre dos veïns es mesura mirant la distància temporal entre les imatges dels extrems que es toquen dels dos veïns.

En la figura 10 podem veure un esdeveniment agrupat pels algorismes de clusterització i divisió-fusió. En aquest esdeveniment podem veure els efectes de la fusió d'un mini-clúster (5 imatges) a un clúster més gran, ja que estava més pròxim temporalment.

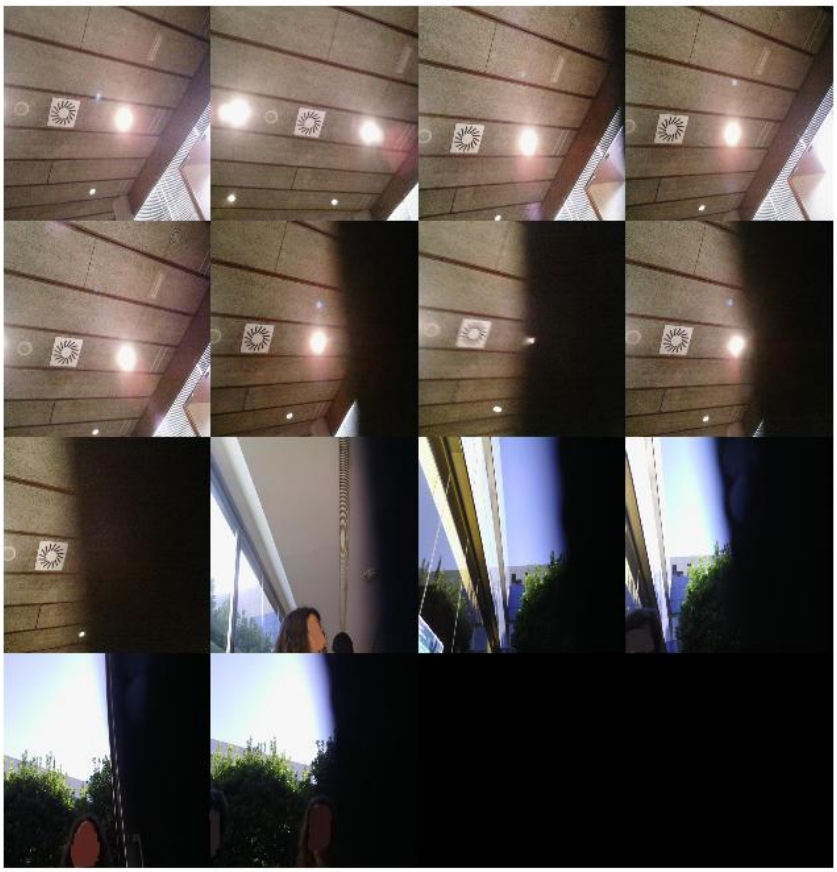


Figura 10. Exemple d'esdeveniment.

4.7. Extracció de l'imatge representativa

Aplicant l'algorisme de divisió-fusió, per fi tenim les imatges del dia separades per esdeveniments. En aquest punt és on apliquem l'extracció de la imatge representativa. Aquí sorgeix la pregunta clau: què vol dir imatge representativa?

Un dels objectius del projecte de la professora Petia Radeva és treballar conjuntament amb psicòlegs professionals per a determinar com contestar precisament aquesta pregunta. En el moment de realització d'aquest treball, encara no s'havia aprofundit en la col·laboració amb els doctors i per tant, amb l'equip vam decidir aplicar el que ens semblava més lògic: l'imatge representativa d'un esdeveniment serà "la més repetitiva" en el sentit de més semblant visualment a la resta d'imatges de l'esdeveniment.

Centrant-nos amb extreure l'imatge més repetitiva, considerem les imatges dels esdeveniments com a nodes de grafs de similitud. Per tant, necessitem un algorisme que assigni a aquests nodes unes puntuacions depenent del nombre d'arestes que rep i els pesos d'aquestes.

4.7.1. Passejada aleatòria

La primera aproximació que vam considerar va ser l'algorisme de Random Walk, o pasejada aleatoria.

Considerem que tenim un vianant a un node del graf i a cada iteració de l'algorisme, aquest vianant ha de moure's a un node veí. Aquest ho farà depenent de la similitud que tingui el node on està amb la resta de nodes, tenint més probabilitat d'anar a un node "més semblant", en el nostre cas, visualment. Això fa que després de cert nombre d'iteracions hi hagi uns nodes per on hagi passat més, i aquests seran els que nosaltres considerarem representatius de l'esdeveniment.

Com s'explica en aquest artícle[5], hi ha una forma ràpida d'obtenir les probabilitats de tots els nodes. Si calculem la matriu de distàncies de l'esdeveniment, és a dir, una matriu on cada fila és la distancia euclidiana d'una imatge respecte a la resta, podem obtenir les puntuacions del graf extraient el principal autovector d'aquesta matriu. Si llavors escollim el major valor, tindrem el l'imatge més representativa de l'esdeveniment.

4.7.2. Distància total mínima

En una reunió del grup, en Marc Bolaños va proposar una altra idea per extreure l'imatge més repetitiva de l'esdeveniment. Considerem que extraiem la matriu de distàncies tal com s'ha descrit al paràgraf anterior. La segona alternativa utilitza aquesta matriu per calcular la "distància total" d'una imatge respecte a la resta. La distància total no és res més que la suma de les distàncies d'una imatge a la resta, per exemple, si tenim que l'imatge A està a una distància de 2 respecte a la B i de 5 respecte a la C, la seva distància total és de 2+5=7. Tenint en compte això, al extreure l'imatge amb menys distància total, estarem extraient l'imatge més repetitiva (o més semblant) de l'esdeveniment. A la figura 11 podem veure dues imatges extretes amb els dos algorismes acabats d'explicar i una tercera escollida aleatòriament.

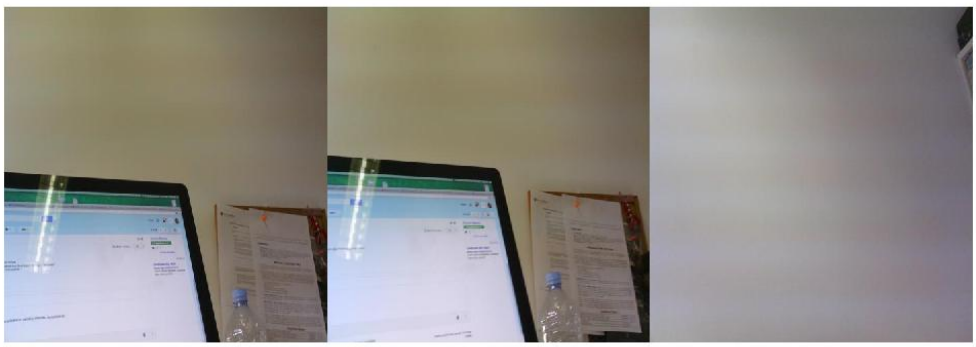


Figura 11. Exemple d'imatges representatives. A l'esquerra passejada aleatòria, al centre distància total minima i a la dreta imatge aleatòria.

Per finalitzar només cal agrupar totes les imatges representatives, una per esdeveniment, per aconseguir el resum visual del dia. A la figura 12 podem veure un exemple de resum visual.

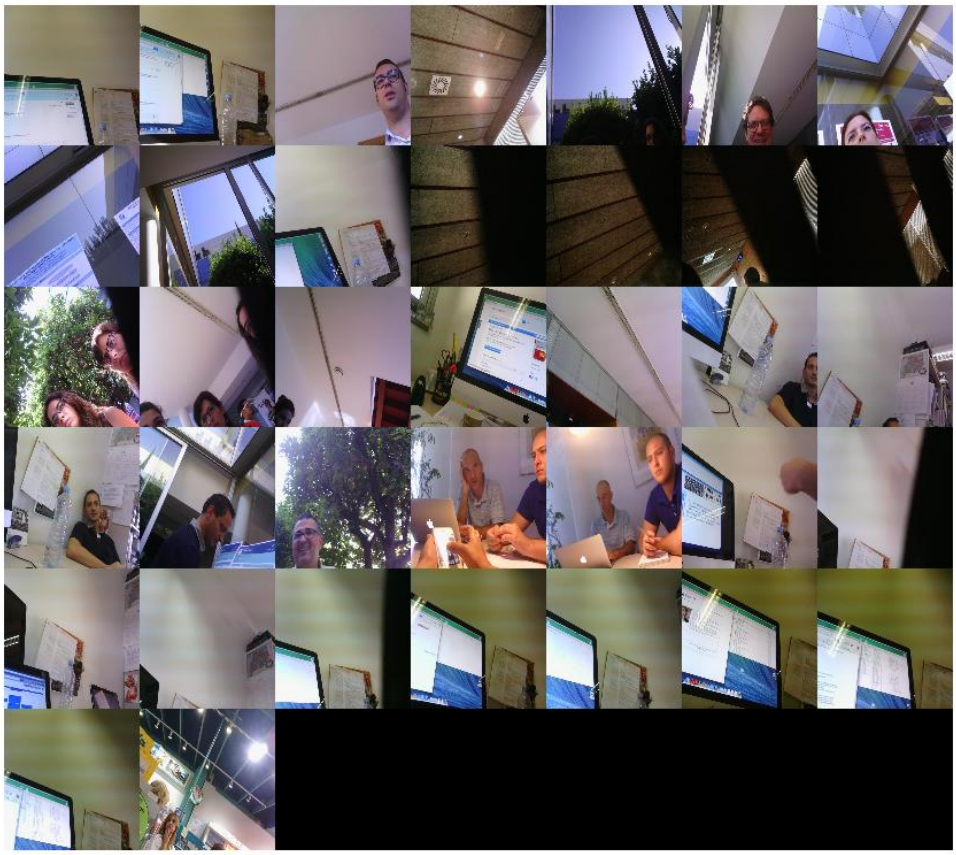


Figura 12. Exemple de resum visual del dia.

5. Resultats

Els resultats finals d'aquest projecte són els resums visuals extrets de cada dia. Tot i així hem volgut avaluar també l'etapa d'agrupament per esdeveniments i l'efecte de l'algorisme de divisió-fusió. Per tant a l'avaluació de resultats hi han dos parts ben definides: una avaluació de la segmentació dels esdeveniments i una avaluació de l'extracció d'imatges representatives.

5.1. Avaluació de la clusterització

En aquesta part de l'avaluació ens preguntem com les diferents variants de l'algorisme d'agrupament i divisió-fusió milloren o empitjoren la segmentació de les imatges en esdeveniments.

Per poder avaluar la segmentació disposavem d'una veritat terreny (o ground truth) per a cada conjunt d'imatges, elaborada pels companys del grup de la UB. Considerant això, vam aplicar l'índex de Jaccard per avaluar la qualitat de la segmentació duta a terme.

5.1.1. Índex de Jaccard

L'índex de Jaccard, també conegut com coeficient de similitud de Jaccard, és una estadística utilitzada per comparar la similitud i la diversitat entre les observacions de dos sets, o conjunts. Es defineix com la mida de l'intersecció dels conjunts dividida per la mida de la unió dels conjunts:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

On:

$$0 \leq J(A,B) \leq 1$$

Equivalent un índex Jaccard de 1 a una similitud perfecta. Cal assenyalar que en el nostre cas l'índex es calcula mirant la intersecció i unió de cada esdeveniment segmentat amb el seu igual de la veritat terreny. Aquesta part del codi va ser proporcionada pel grup BCNPCL de l'UB, ja que la tenien implementada per avaluar els seus experiments.

Un cop definida la nostra mesura d'avaluació, presentarem les proves que hem realitzat amb les diferents variants de l'algorisme de clusterització i divisió-fusió. Compararem resultats amb la millor configuració que hem obtingut, és a dir, la configuració que ha comportat un índex de Jaccard més gran: característiques normalitzades, enllaç *average* i divisió-fusió temporal. També cal destacar que l'agrupament depèn principalment del paràmetre *cutoff*, com ja s'ha explicat a la secció de metodologia. Per tant hem efectuat les proves per a diferents valors d'aquest paràmetre. Totes les avaluacions que es presenten a continuació s'han fet amb la mitjana dels cinc índexs de Jaccard als cinc conjunts de fotografies.

5.1.2. Efecte dels métodes d'enllaç

Per començar justificarem l'elecció del mètode d'enllaç *average*. Ho hem fet provant tots els métodes d'enllaç dels que disposem a la funció *clusterdata* de Matlab:

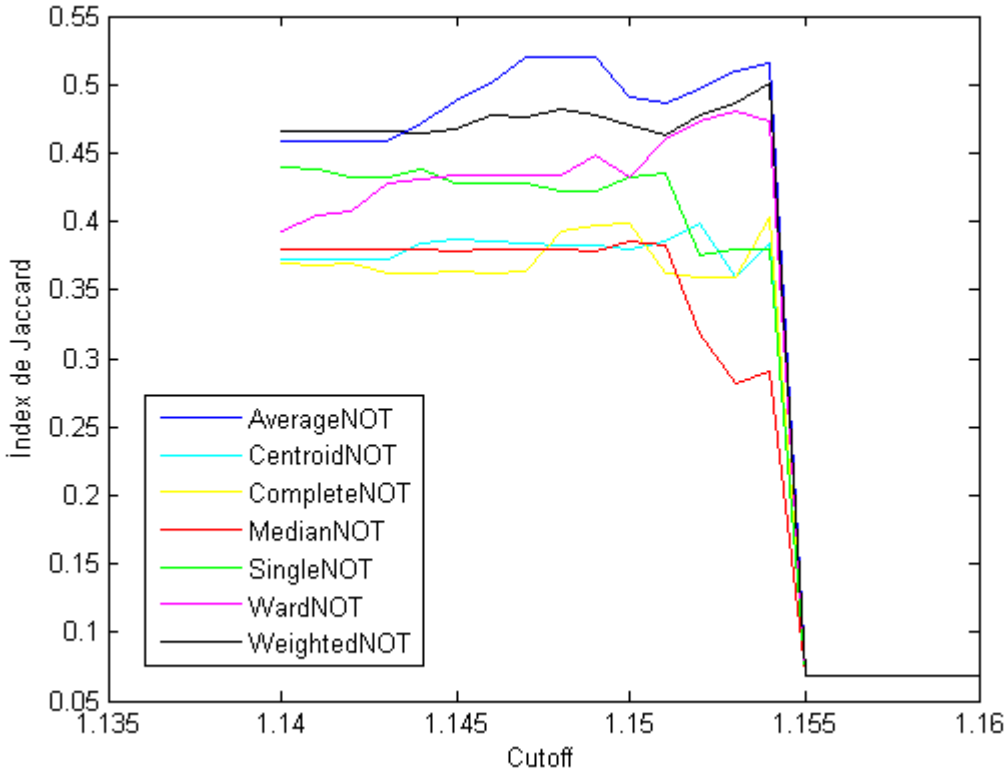


Figura 13. Índex de Jaccard, comparació dels mètodes d'enllaç.

La nomenclatura és la següent: el *NOT* darrere de cada métode significa Normalitzat i Ordenat Temporalment. Com ja s'ha comentat anteriorment, podem observar que el millor métode d'enllaç és l'*average*, tot i que seguit de prop per el métode *weighted* i el *ward*, el que pot significar que en altres sets hi hagi la possibilitat que funcionin millor que l'*average*. També podem deduir que els clústers d'imatges són tenen una distribució prou uniforme ja que la mitjana de totes les distàncies acaba sent la millor alhora de classificar. Ens hem quedat amb el métode *average* ja que és computacionalment més ràpid.

5.1.3. Comportament general

També cal observar el comportament general de l'índex depenent dels valors de cutoff que fan que l'agrupament sigui no supervisat (entre 0 i 2). En aquest cas aplicant normalització de característiques, clusterització *average* i divisió-fusió temporal:

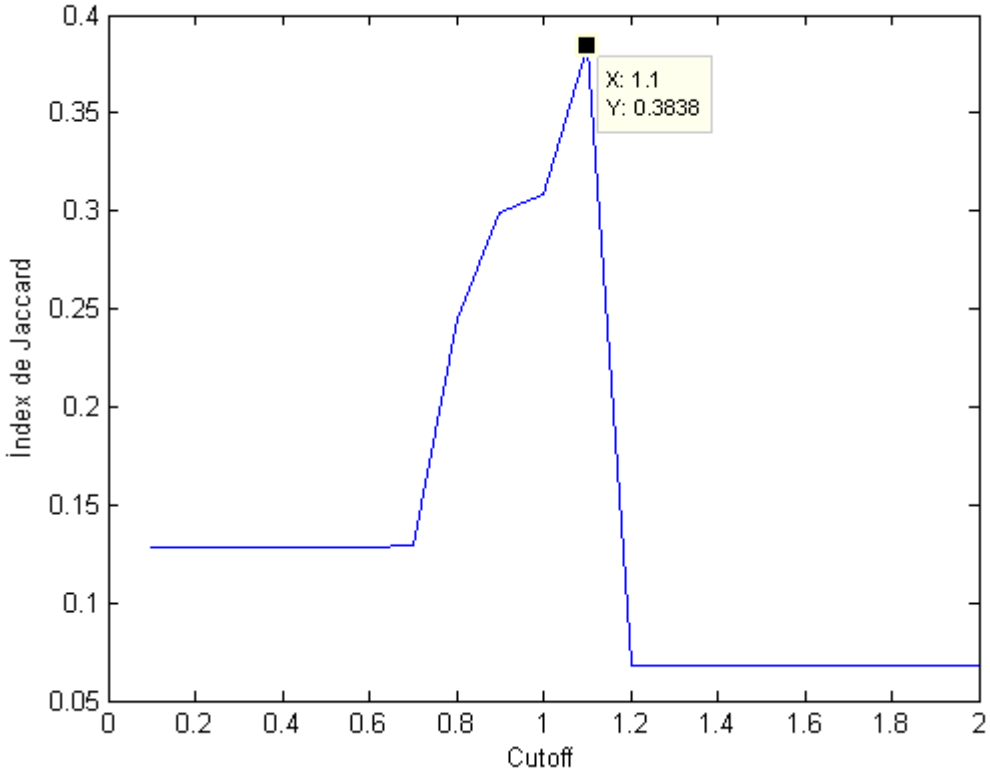


Figura 14. Índex de Jaccard general.

Com es pot observar hi ha un rang de valors on podem trobar l'òptim, i és on hem fet la resta d'avaluacions i també l'anterior: l'interval de *cutoff* entre 1.14 i 1.16.

5.1.4. Efecte de la normalització

Com ja s'ha explicat, la normalització dels descriptors visuals ajuda a classificar més eficientment ja que tots els valors estan en el mateix rang:

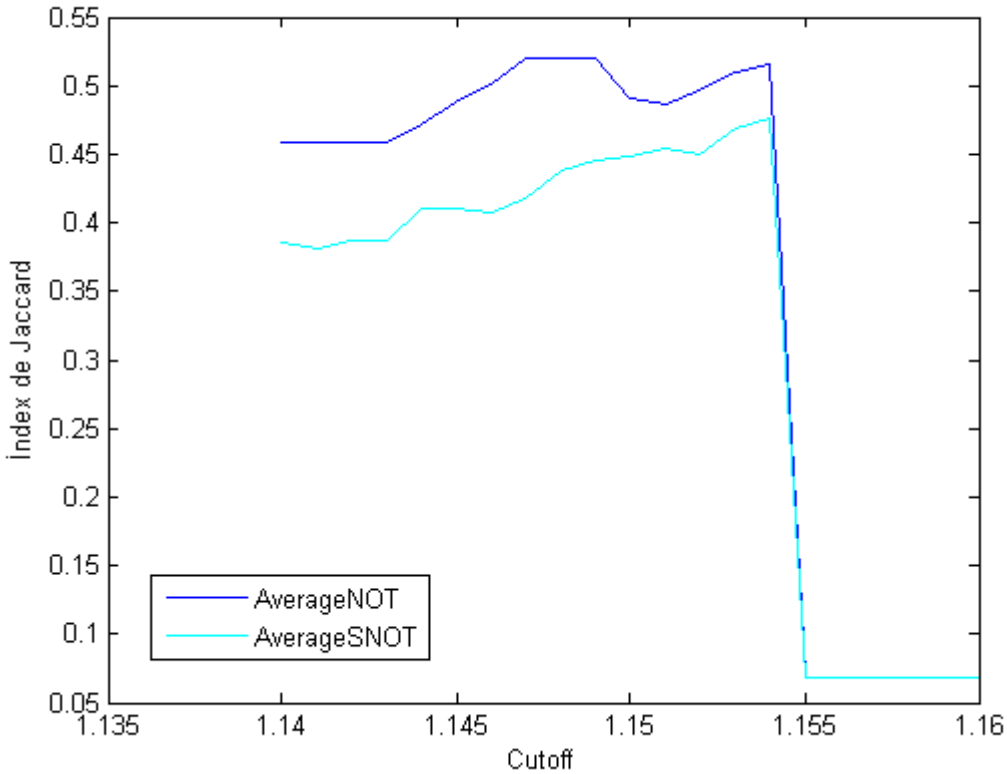


Figura 15. Índex de Jaccard, característiques normalitzades i no normalitzades.

On *AverageNOT* vol dir normalitzat i ordenat temporalment, i *AverageSNOT* vol dir Sense Normalitzar i ordenat temporalment. Podem observar que el fet de normalitzar les característiques, realment millora el rendiment de la clusterització.

5.1.5. Efecte de la divisió-fusió

Un cop vistos els efectes de normalitzar i escollir diferents métodes d'enllaç, ara cal avaluar l'efecte de l'algorisme de divisió-fusió que hem dissenyat. Aquest algorisme, d'una banda divideix els esdeveniments més grans, si entremig hi ha un esdeveniment de 6 o més imatges, per poder adaptar-nos millor a la veritat terreny. D'altra banda l'algorisme també reetiqueta els grups de menys de 5 imatges a esdeveniments més grans.

Com ja hem dit a la secció de metodologia, considerem que no hi han esdeveniments de cinc o menys imatges per la seva curta durada temporal. En la veritat terreny vam detectar uns quants d'aquests esdeveniments. Per ser coherents amb l'avaluació, vam eliminar d'aquesta avaluació les imatges pertanyents a aquests petits esdeveniments, el qual vol dir que no les vam incloure a l'hora de classificar-les ni a l'hora de mirar la veritat terreny per extreure l'índex de Jaccard.

La primera gràfica explora la possibilitat de no aplicar cap reetiquetatge temporal ni considerar que els esdeveniments de cinc o menys imatges no són representatius:

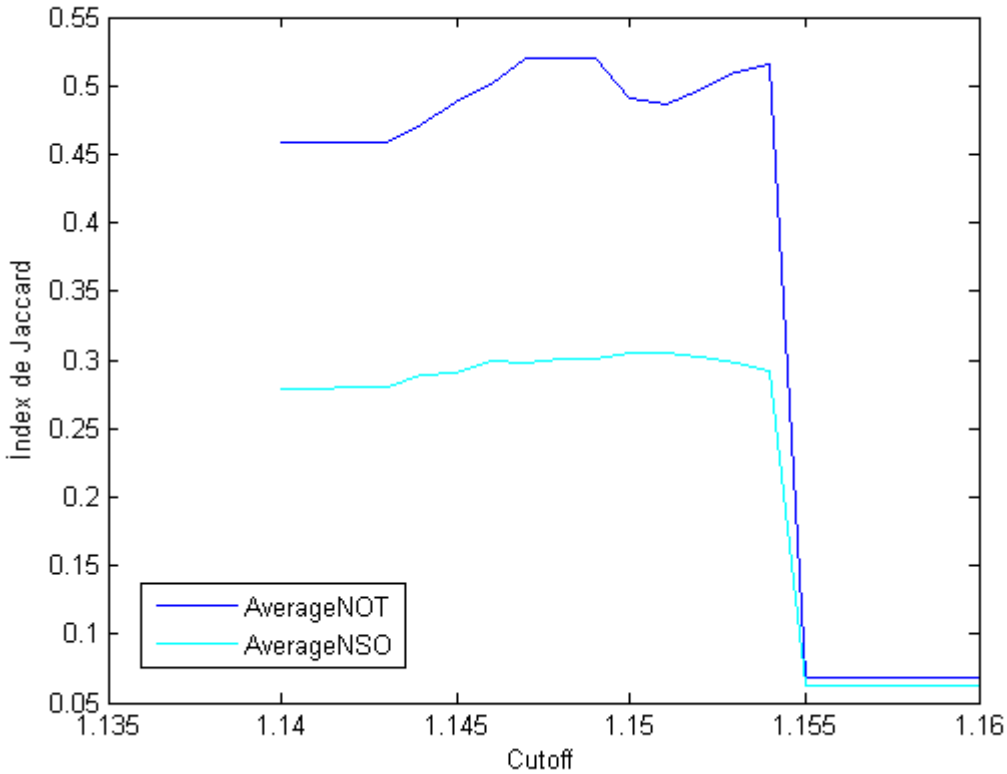


Figura 16. Índex de Jaccard, ordenat i sense ordenar.

Pel que fa a la nomenclatura nova, “NSO” significa Normalitzat i Sense Ordenar. En aquesta gràfica podem veure que realment aplicar l’algorisme de divisió-fusió és beneficiós, ja que practicament dobla l’índex de Jaccard que obtenim si no ordenem i no descartem esdeveniments petits (5 o menys imatges).

En la següent gràfica comparem la primera alternativa que vam utilitzar, assignar els esdeveniments petits a l’esdeveniment anterior més gran, i l’alternativa de assignar-los depenent de la proximitat temporal dels esdeveniments anterior i posterior:

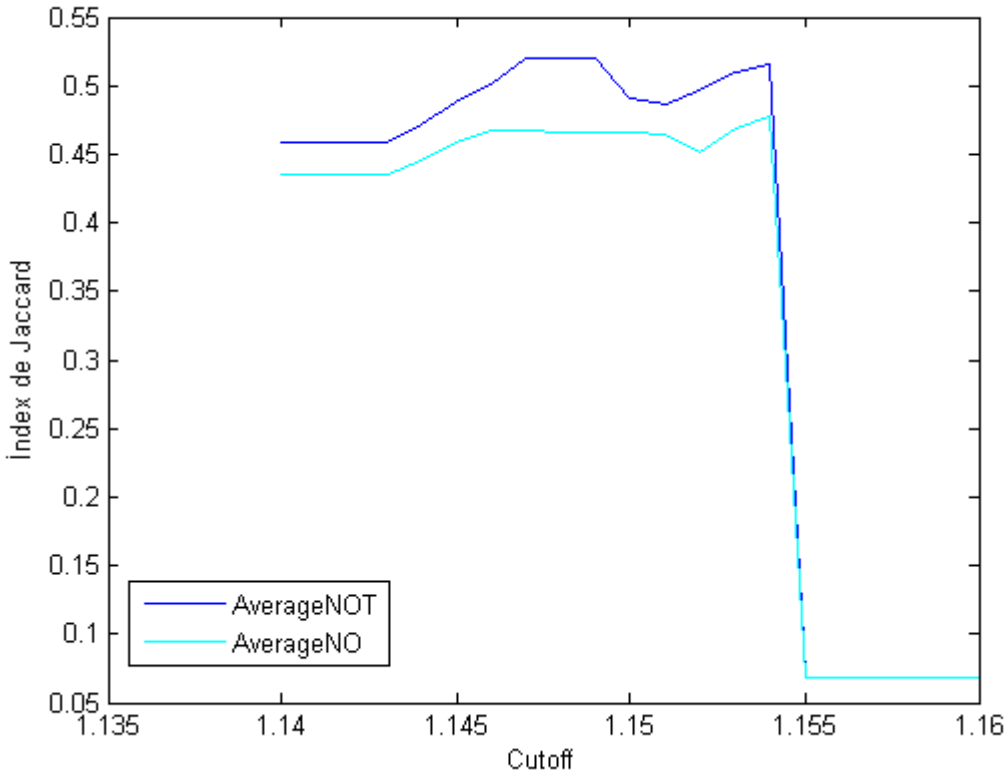


Figura 17. Índex de Jaccard, ordenat (no temporalment) i ordenat temporalment.

"NO" significa normalitzat i ordenat (no temporalment). Podem veure que el fet de tenir en compte l'informació temporal proporcionada per Narrative fa millorar sensiblement el rendiment de l'algorisme.

Com ja s'ha avançat, la millor configuració per a l'agrupament i divisió-fusió resulta ser característiques normalitzades, mètode d'enllaç *average*, reetiquetatge temporal i valor experimental de cutoff=1.154. Utilitzant aquesta configuració hem extret els resums visuals del següent apartat.

5.2. Avaluació de l'extracció d'imatge representativa

A diferència de l'agrupament, en aquesta etapa no tenim una veritat terreny sobre la que recolzar-nos. Per tant per avaluar els resultats de l'elecció de les imatges representatives i de la qualitat del resum visual obtingut, ens hem volgut basar en el mateix mètode que fa servir Kristen Grauman al seu treball[4].

La técnica es denomina "tast a cegues" (en anglès, "blind taste") i per al nostre cas l'hem aplicat de la següent forma. Hem separat l'avaluació en dues parts, representativitat de l'imatge triada dins de l'esdeveniment i qualitat del resum visual obtingut.

S'ha utilitzat l'eina "Google forms"[8] per recollir les opinions d'un total de 20 persones. Al principi del qüestionari, s'explica el funcionament del mateix i és demana a l'enquestat que no tingui en compte si la segmentació temporal és l'òptima o no, ja que aquest aspecte s'avalua d'una altra forma, com hem vist a l'apartat anterior.

5.2.1. Avaluació de la representativitat individual

En la primera part del qüestionari es presenta a l'enquestat cada esdeveniment segmentat automàticament, és a dir, un cop processades les imatges per l'algorisme de clusterització i pel de divisió-fusió, en forma de matriu d'imatges:

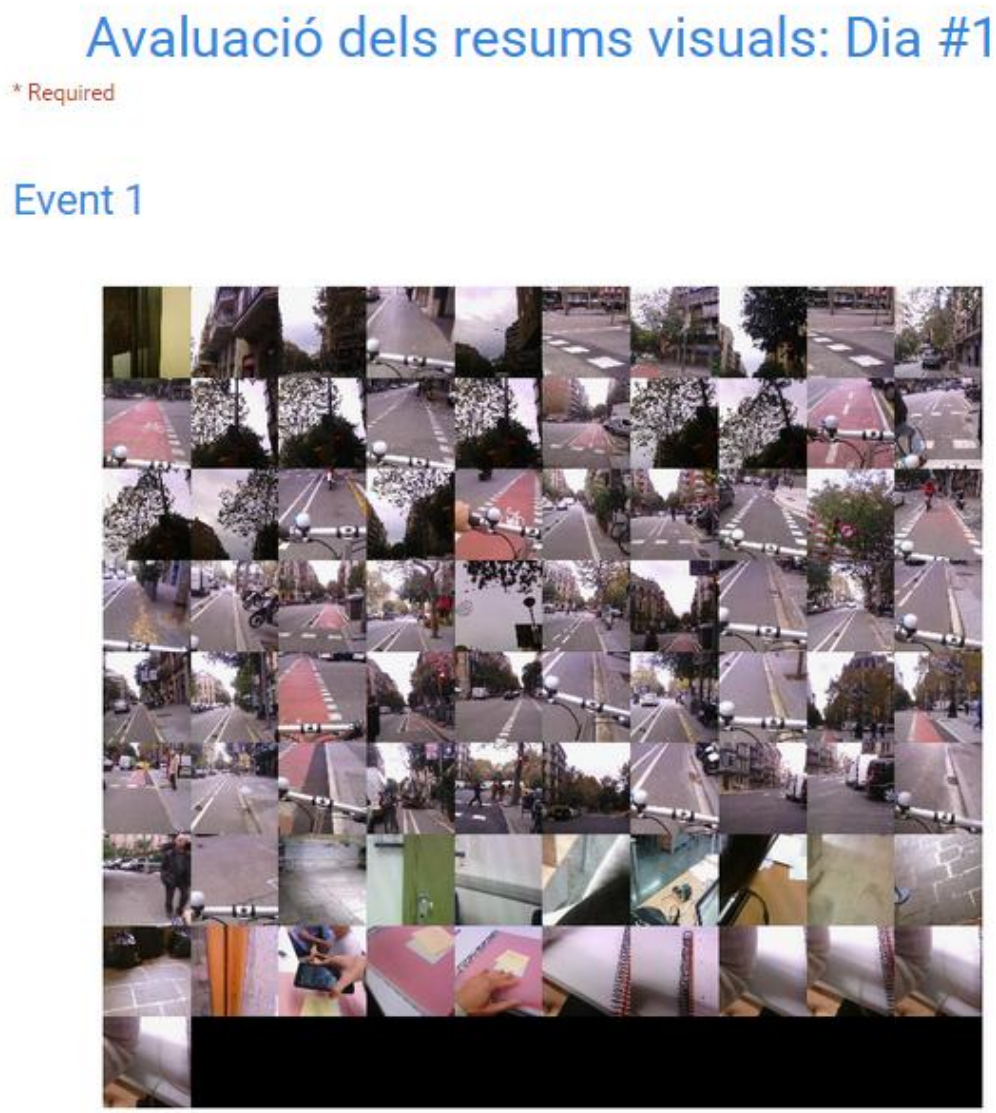


Figura 18. Exemple de presentació d'un esdeveniment.

Així aconseguim que l'usuari és fa una idea de que ha passat a l'esdeveniment. Aquest fet també ens serà útil de cara a la segona part de l'avaluació. Seguidament se li presenten tres imatges representatives de l'esdeveniment (ordenades de manera aleatòria a cada pàgina): una escollida amb la técnica passejada aleatòria, una altra amb

la técnica de la distància total mínima, i una tercera escollida aleatòriament. La raó de presentar una imatge aleatòria és per comprovar si el fet de extreure una imatge amb les tècniques anteriors millora la técnica més bàsica de totes, que és escollir una imatge aleatòriament.

Cal destacar, que en la gran majoria de casos (86% dels esdeveniments) les tècniques de distància total mínima i passejada aleatòria coincideixen en la imatge escollida. Degut a això, en els esdeveniments que passa, només presentem dues imatges, la escollida per les tècniques i la escollida aleatòriament per evitar redundància als resultats.

De les tres (o dues) preguntem si poden representar l'esdeveniment, i posteriorment quina triaria com a millor representativa de les tres (o dues). Amb això aconseguim avaluar la representativitat de cada imatge i comparar entre elles la qualitat de la representativitat. A cada esdeveniment les presentem ordenades diferent, per evitar cap tendència repetitiva:

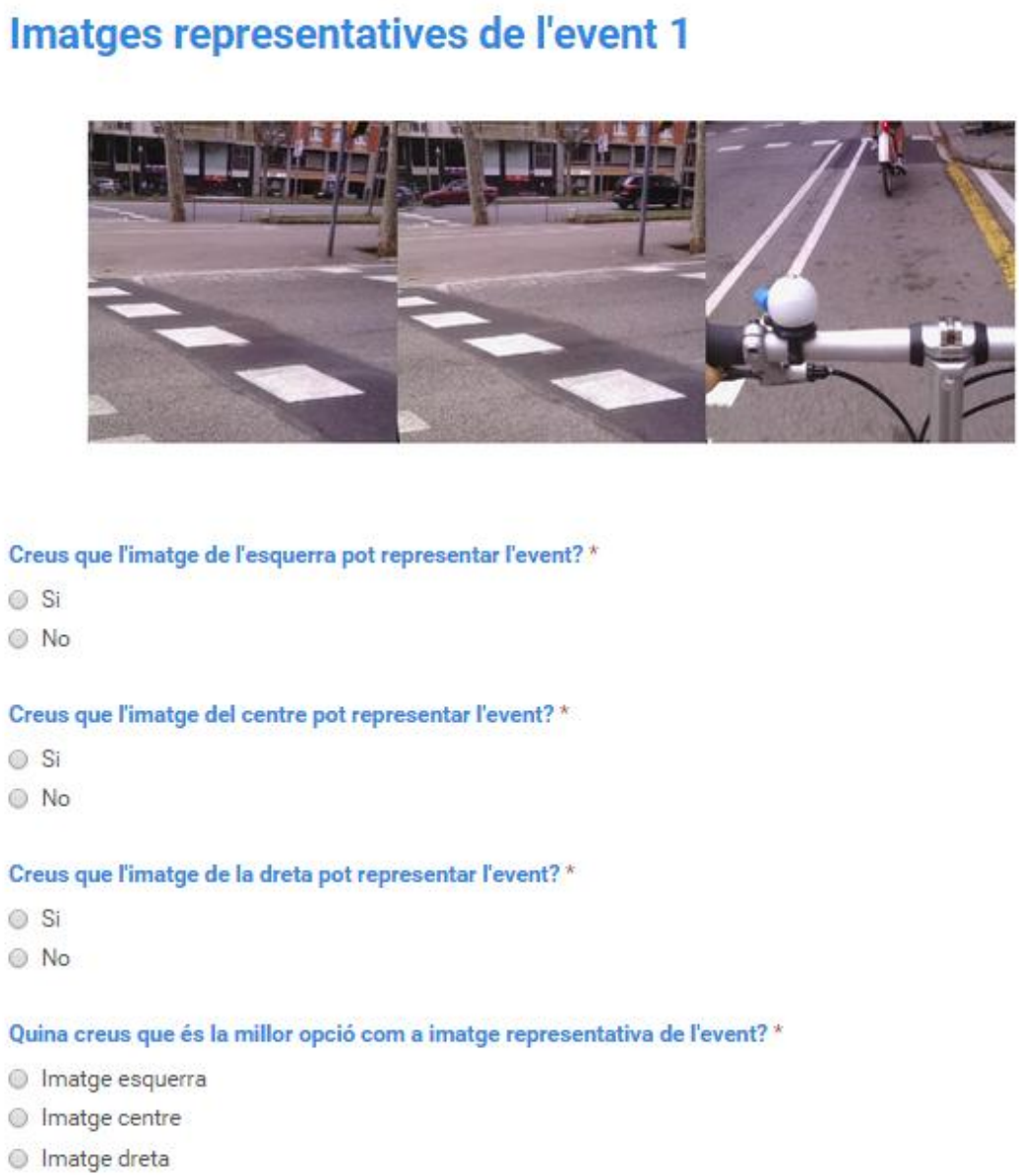


Figura 19. Exemple de presentació de les imatges representatives de l'esdeveniment.

Un cop explicat com s'han recollit les opinions sobre la representativitat individual, passem a intentar explicar els resultats obtinguts:

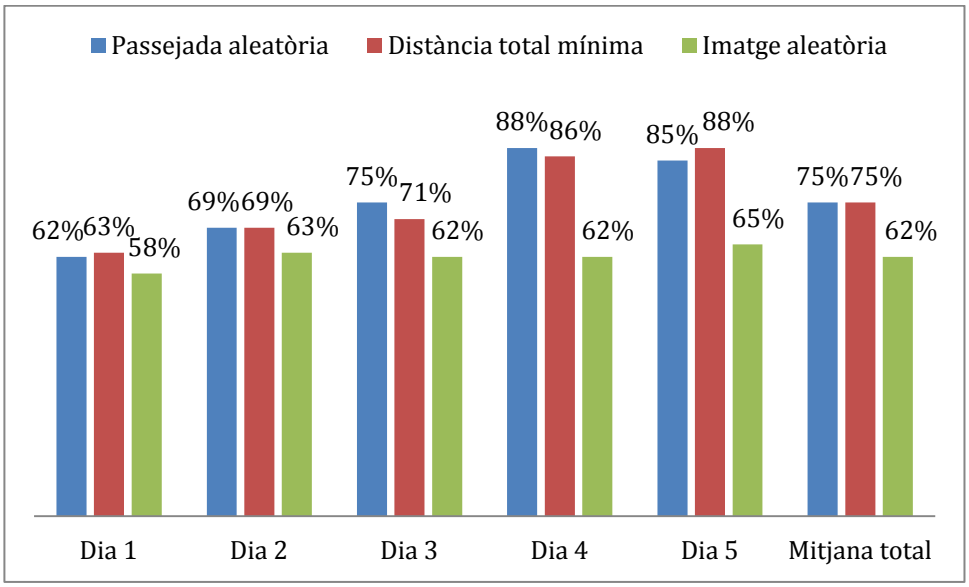


Figura 20. Validacions de les tècniques, individualment.

Com ja s'ha explicat, per a fer les avaluacions comptavem amb cinc conjunts d'imatges per cinc dies de captura amb *Narrative*. A la primera columna veiem el percentatge de enquestats que han opinat que l'imatge extreta mitjançant la técnica passejada aleatòria pot representar l'esdeveniment en el qual se'ls preguntava. Podem interpretar les dues columnes restants igualment.

Com veiem les imatges extretes amb les tècniques elaborades en aquest treball han estat més votades que l'imatge aleatòria. Per aprofundir en aquest fet hem de pensar com són els esdeveniments que s'han segmentat a l'etapa de clusterització i divisió-fusió. Com hem vist, el segmentat no és ni molt menys perfecte, la qual cosa provoca que alguns dels esdeveniments segmentats automàticament continguin imatges que es poden considerar d'un altre esdeveniment:

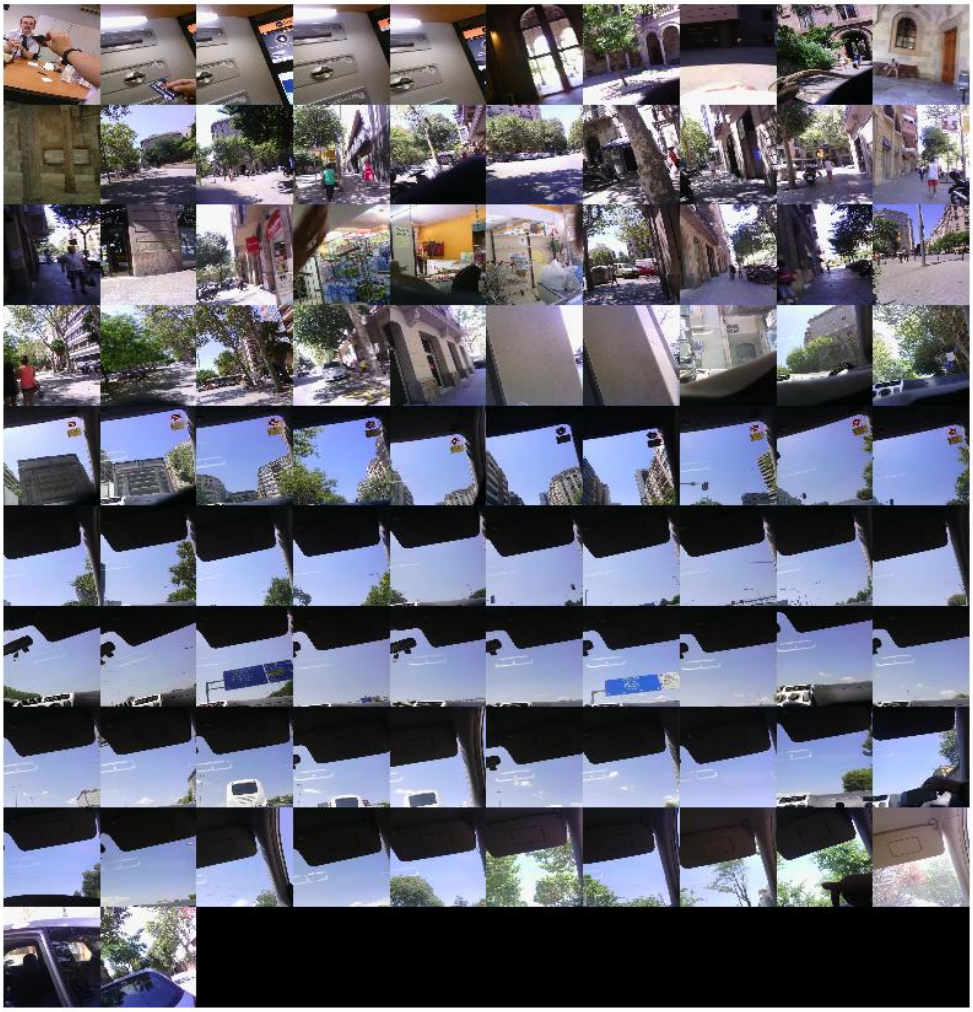


Figura 21. Exemple d'un esdeveniment automàtic amb imatges de dos esdeveniments.

En aquest esdeveniment les primeres imatges són d'una passejada pel carrer i les següents (i majoritàries) són d'un viatge amb cotxe.

Imaginem ara que la segmentació ha estat perfecta, obtenint un índex de Jaccard de 1. Els esdeveniments estarien perfectament separats, i per tant les imatges de dins serien extremadament semblants. Si aquí hi apliquem les tècniques de passejada aleatòria i distància total mínima, obtindrem l'imatge representativa visualment, que serà la més semblant a la resta. Si ara escollim una imatge aleatoriament també tindrem una imatge molt semblant a la resta i si l'usuari no sap aquest fet, la podria confondre perfectament amb una imatge escollida per tècniques més elaborades. Això provocaria que els resultats fossin molt igualats ja que al ser imatges pràcticament iguals, la subjectivitat de l'usuari igualaria resultats.

Com que la segmentació dels esdeveniments no ha estat perfecta podem tenir esdeveniments com l'anterior, que són una mescla de dos i fins i tot tres esdeveniments. En aquest cas les tècniques de passejada aleatòria i distància total mínima extrauran la més semblant, i com la majoria (però no totes) d'imatges són de l'esdeveniment en concret, n'escolliran una de les de la majoria. En canvi l'escollida aleatoriament pot ser

una de les mateixes o una dels esdeveniments *incrustats* a l'esdeveniment més gran. Això farà que la puntuació de l'imatge aleatòria disminueixi, com es pot constatar als resultats.

Tot aquest raonament també es posa de manifest a la següent figura, que són els resultats a la pregunta de quina imatge representativa és la millor:

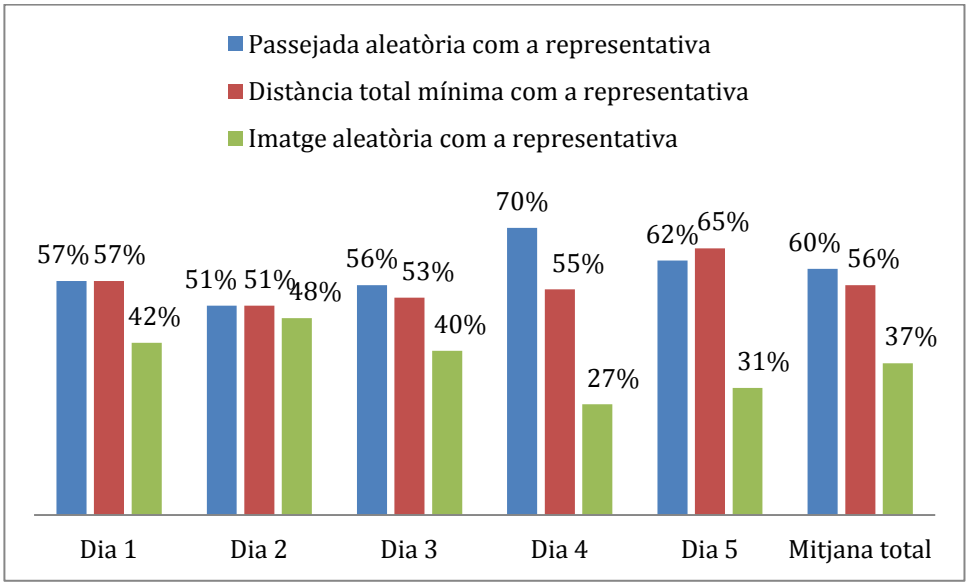


Figura 22. Tria de tècniques com a representativa.

Un primer cop d'ull podria fer-nos pensar que els resultats d'aquesta figura haurien de ser complementaris (per exemple 33% per passejada aleatòria, 33% per a distància total mínima i 33% per a l'imatge aleatòria), però no em oblidar el fet de que el 85% de les imatges representatives extretes amb passejada aleatòria i distància mínima total coincideixen. Per tant, per cada vot que rebia una d'aquestes imatges es puntuava a les dues tècniques. Tot i així si que podem mirar la complementarietat de resultats tenint en compte que el 37% de les vegades l'enquestat a escollit l'aleatòria com a representativa, fet que implica que el 63% de les vegades no l'ha escollit i per tant ha escollit una de les altres tècniques.

5.2.2. Avaluació de la representativitat conjunta

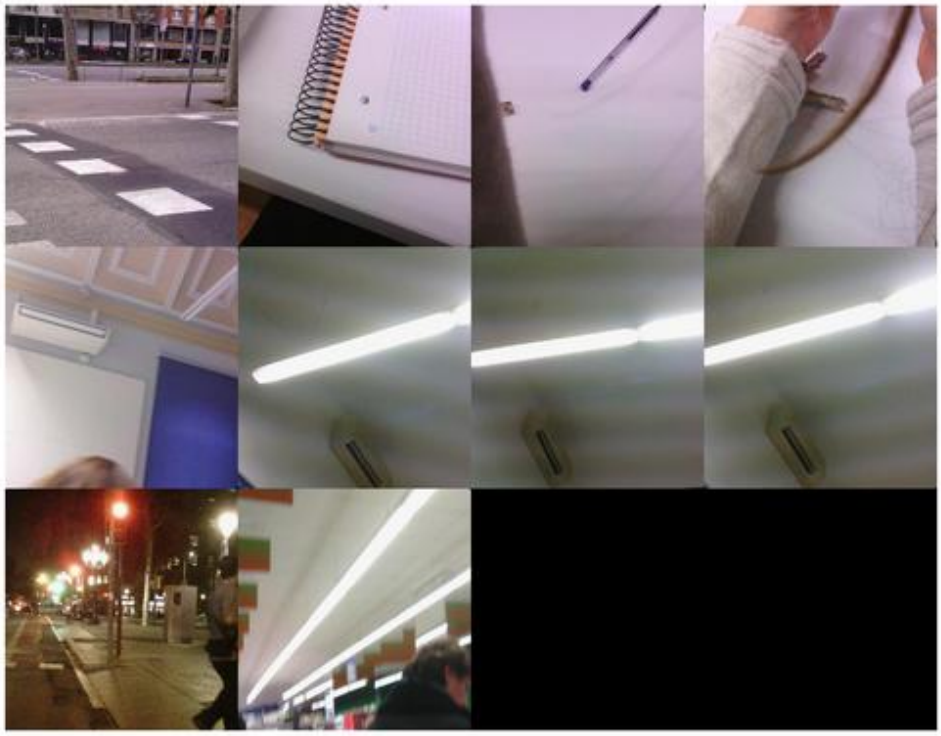
Un cop avaluada l'elecció de la imatge representativa, hem volgut avaluar també el conjunt del resum visual final. Per fer-ho, a l'última pàgina del qüestionari hem presentat quatre resums visuals: el primer obtingut amb totes les imatges extretes per passejada aleatòria, el segon obtingut amb la técnica de distància total mínima, el tercer amb les imatges aleatòries escollides anteriorment i finalment un conjunt d'imatges escollides uniformement en el temps durant tota la col·lecció.

Per cada resum, preguntem si l'enquestat creu que podria servir com a resum visual del dia. Amb aquesta pregunta pretenem avaluar la qualitat general del resum visual, i per tant, la qualitat més en general de la representativitat extreta per les diferents tècniques:

Resums visuals del dia

Pot ser que vegem dos resums molt semblants (es diferenciaran amb poques imatges). En aquest cas podem triar com a millor qualsevol dels dos.

Resum 1



Creus que aquest conjunt d'imatges pot resumir el dia complet? *

- Si
- No

Figura 23. Exemple de presentació d'un resum visual complet.

I finalment se li pregunta per quin considera que és el millor dels quatre resums presentats:

Finalment, quin creus que és el millor resum visual del dia? *

- Resum 1
- Resum 2
- Resum 3
- Resum 4

Figura 24. Pregunta final.

Un cop presentada la forma en que hem preguntat als enquestats, volem analitzar la representativitat conjunta de cada tècnica, avaluant el conjunt d'elles com un resum visual del dia:

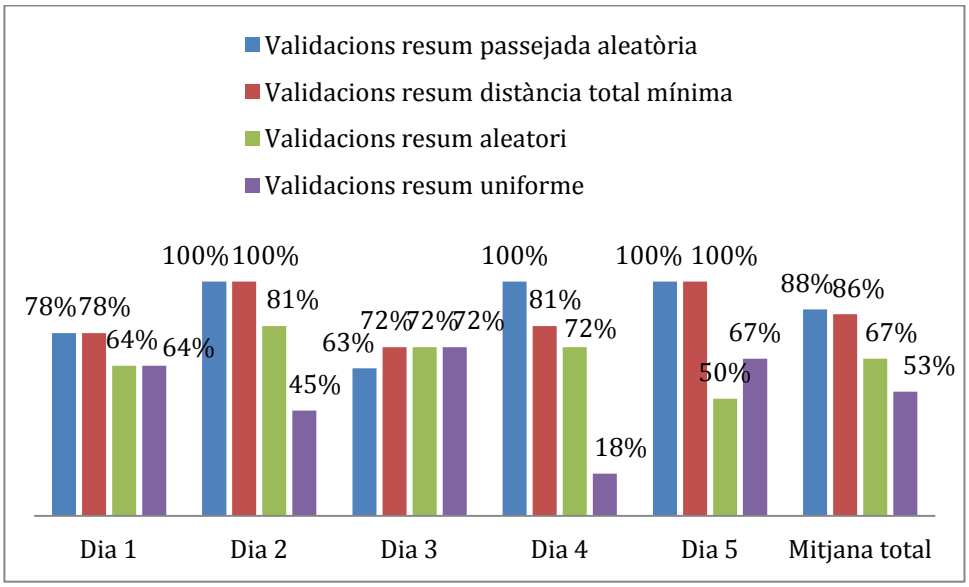


Figura 25. Validacions de les tècniques com a resums visuals.

La primera columna de la taula representa el percentatge d'usuaris que han opinat que el conjunt d'imatges representatives extretes amb passejada aleatòria poden resumir el dia. Com anteriorment han estat puntuant els diferents esdeveniments del dia, el factor de no recordar molt bé com ha sigut el dia no influeix. El mateix es pot aplicar a la resta de columnes.

Com veiem, les puntuacions de passejada aleatòria i distància mínima total són superiors a les obtingudes amb el resum aleatori i l'uniforme. Aquest fet es deu a l'explicat anteriorment: les imatges escollides aleatoriament poden no contenir representativitat de certs esdeveniments i això penalitza la seva elecció. Per explicar la baixada del resum uniforme també ens podem valer del mateix raonament, ja que en aquesta elecció no es té en compte la segmentació dels esdeveniments ni la representativitat visual, simplement són imatges escollides equiespaiades temporalment.

A l'última taula podem veure les puntuacions a la pregunta de quin consideren els enquestats el millor resum visual del dia que se'ls ha presentat:

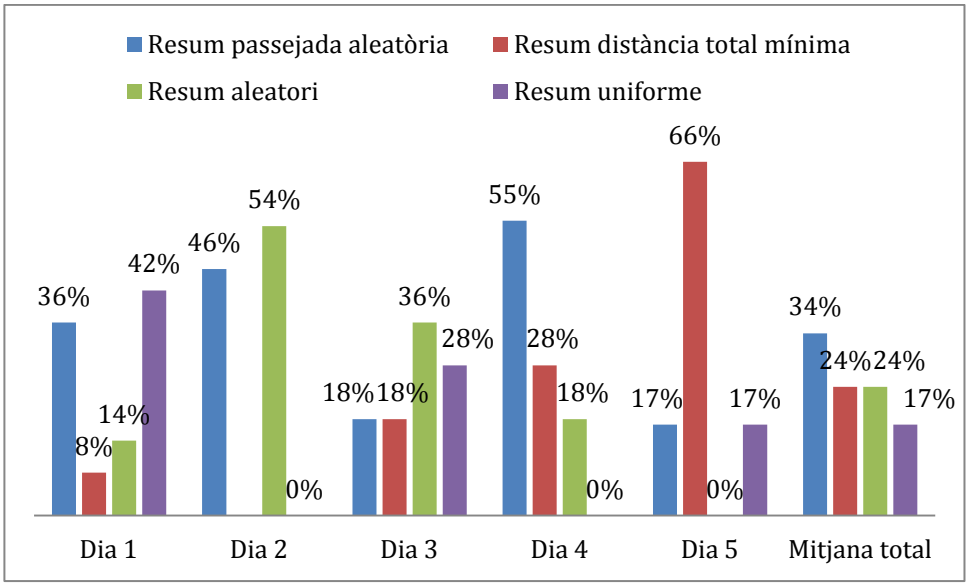


Figura 26. Tria del millor resum visual.

Ara si que els resultats són sempre complementaris perquè els resums visuals extrets amb passejada aleatòria i distància total mínima no coincidien més que en un cop, i per tant el resum de distància total mínima no es va presentar per avaluar.

També cal senyalar que els resums amb les tècniques dissenyades en aquest treball eren molt similars, canviant una o dues imatges l'un de l'altre. Aquest fet ha provocat que es repartissin entre si les puntuacions depenent de la subjectivitat de l'usuari. Si sumem les seves puntuacions veiem que un 58% de les vegades l'usuari ha opinat que el millor resum era un dels elaborats amb tècniques que tenen en compte la similitud visual de les imatges.

6. Pressupost

Aquest projecte s'ha dut a terme amb el programari Matlab exclusivament, amb les llicències del grup d'imatge de l'UPC. Per tant com a costos tindrem: el cost de la llicència de Matlab i el cost del personal associat.

Agafant com a guía el sou que guanya un becari de la Universitat Politécnica de Catalunya i el total d'hores dedicades al projecte tenim:

	Personal	Preu/Hora	Dedicació	Sou	#Mesos	Total Personal
Enginyer junior	1	8.00 €/hora	80 hores/mes	640 €	5	3200 €

Taula 1. Cost del personal associat.

Programari utilitzat	Unitats	Preu/Unitat	Total Programari
Matlab	1	500 €	500 €

Taula 2. Cost de la llicència del Matlab.

Cost total del projecte
3700 €

Taula 3. Cost total.

7. Conclusions i treball futur:

7.1. Conclusions

L'objectiu principal d'aquest treball era extreure un resum visual del dia, tenint en compte la similitud visual de les imatges i les seves marques temporals. Com hem pogut veure a l'apartat de resultats l'objectiu s'ha complit i amb bones puntuacions per part dels usuaris enquestats superiors a solucions que no utilitzessin cap processament del senyal. Com a avaluació general del sistema ens podem fixar amb la figura 26 on es veu plasmada l'opinió sobre la qualitat dels resums extrets. S'han obtingut un 58% d'opinions favorables a resums extrets amb tècniques elaborades en aquest treball: passejada aleatòria i distància total mínima.

També hem pogut comprovar com apart de l'informació visual de les imatges, pot ser molt útil considerar també l'informació temporal per agrupar per esdeveniments les imatges, o simplement per reetiquetar imatges que pertanyen a esdeveniments petits (cinc o menys) i que no els considerem esdeveniments per la seva curta durada. Tenint en compte l'informació temporal hem pogut obtenir un índex de Jaccard del 0.5, mentre que considerant altres criteris només hem pogut obtenir 0.45 per al reetiquetage sense tenir en compte les marques temporals i 0.3 sense cap reetiquetatge.

Com a conclusió final podem dir que aquest treball, tot i no obtenir resultats molt alts, representa una base per a treball futur sobre com resumir dies tenint en compte similitud visual, i com veurem al treball futur, diferents criteris semàntics sobre l'elecció d'una imatge representativa. Un altre estudiant de l'UPC proseguirà el treball a partir d'aquest pròxim quadrimestre de primavera de 2015.

El codi elaborat i utilitzat, que s'explica en l'annex (secció 9) en aquest projecte es publicarà a la web del GPI. També cal destacar que s'espera enviar un artícle científic basat en aquest projecte a 11th IEEE Embedded Vision Workshop del CVPR 2015 [15].

7.2. Treball futur

Amb l'objectiu de millorar els resultats obtinguts en aquest treball es plantegen les següents possibles millores.

En primer lloc, per a millorar l'etapa de clusterització es proposa provar diferents algoritmes d'agrupament com *mean shift* o *spectral clustering*. Es volien provar en aquest treball però degut a la falta de temps no va ser possible.

En segon lloc l'opció de aplicar un detector d'objectes, com es volia fer al principi, donaria un sentit semàntic als resultats. En la proposta actual només es té en compte les característiques visuals de les imatges per a extreure l'imatge representativa d'un esdeveniment. Aquest fet comporta que alguns cops l'imatge representativa sigui la més semblant visualment a la resta, però no contingui res que es consideri important, com persones o objectes, que poden implicar l'importància del moment.

Aplicant detectors d'objectes i de cares, podriem obtenir uns resultats semànticament més rics, ja que contindrien persones i objectes, que al cap i a la fi, són els que doten de importància els moments del dia.

Per últim, es va consultar a uns experts psicòlegs, de quina forma podem considerar representativa una imatge per al dia. En aquest treball només s'ha tingut en compte la imatge més representativa visualment, per tant s'ha agafat la més "repetitiva" de l'esdeveniment. En canvi els experts aporten més alternatives per escollir representativitat i que en un treball futur es poden tenir en compte per millorar els resums:

-Imatge més repetitiva (utilitzat en aquest treball).

-Imatge poc usual dins de l'esdeveniment. Que podem traduir com la menys repetitiva. Aquesta opció es basa en que acabem recordant o considerant importants els moments menys repetitius o poc usuals del dia.

-Imatge que contigui informació de l'entorn. Falta concretar com, però pot ser una informació important ja que no tots els dies es va als mateixos llocs, i això pot ser representatiu del dia.

-Imatge que contingui interaccions socials. A la proposta anterior ja s'ha comentat com pot millorar els resultats el fet de tenir en compte les persones.

-Imatge representant una activitat. També queda pendent concretar com es pot extreure aquesta informació.

-Imatge que contingui informació de quan s'ha produït un esdeveniment.

-Imatge que expliqui com s'ha produït un esdeveniment o activitat.

8. Bibliografía:

[1] D. Almendros, “Visual instance mining of news videos using a graph-based approach” de *Departament de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya*.

[2] Giro-i-Nieto, Xavier, Monica Alfaro, and Ferran Marques. "Diversity ranking for video retrieval from a broadcaster archive." *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM, 2011.

[3] V. Chandrasekhar, W. Min, X. Li, C. Tan, B. Mandal, L. Li, J. H. Lim, “Efficient Retrieval from Large-Scale Egocentric Visual Data Using a Sparse Graph Representation” de *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference*.

[4] Z. Lu and K. Grauman, “Story-driven summarization for egocentric video,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[5] X. Giró, “A random walk, step by step” de *bitsearch.blogspot.com.es*

[6] *http://dl.acm.org/citation.cfm?doid=2647868.2654889*

[7] “Base de dades visual ImageNet” de *http://www.image-net.org/*

[8] “Google Forms” de *http://www.google.com/forms/about/*

[9] M. Martos, “Content-based video summarization in Object Maps” de *Departament de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya*.

[10] A. R. Doherty, D. Byrne, A. F. Smeaton, G. J. Jones, and M. Hughes, “Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs,” *in Proceedings of the 2008 international conference on Content-based image and video retrieval. ACM, 2008, pp. 259–268*.

[11] Y. J. Lee, J. Ghosh, and K. Grauman, “Discovering important people and objects for egocentric video summarization,” *in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 1346–1353*.

[12] O. Aghazadeh, J. Sullivan, and S. Carlsson, “Novelty detection from an ego-centric perspective,” *in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011, pp. 3297–3304*.

[13] A. Fathi, J. K. Hodgins, and J. M. Rehg, “Social interactions: A first-person perspective,” *in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 1226– 1233*.

[14] M. Bolaños, M. Garolera and P. Radeva, “Video segmentation of life-logging videos”

[15] http://cvisioncentral.com/promotion/evw2015/

[16] Ventura, C., Giro-i-Nieto, X., Vilaplana, V., Giribet, D., & Carasusan, E. (2013, June). Automatic keyframe selection based on mutual reinforcement algorithm. In Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on (pp. 29-34). IEEE.

9. Annexos: Developer guide

This section will explain briefly how the project code works. There are two main sets of functions, on one hand the functions that compute the visual summary of a set of images, and on the other hand the ones that compute the evaluation of different project parameters.

9.1. Visual summary extraction functions

9.1.1. mainVisualSummary

This function extracts the visual summary of all the images of the specific path.

Inputs:

-path: string containing the path of the images to summarize. Example of path: path='/imatge/rmestre/work/Images/testImages/Petia2';

-montageFlag: boolean. If it's value is true, the code will plot every event and it's representative images. This is useful if we want to extract figures with the complete events and the representative images to evaluate the results. It is recommended to put a breakpoint at the "montage(im)" line, if not all figures will overwrite the previous one.

Output:

-visualSummaryRandomWalk: a cell array containing the paths of the representative images extracted through random walk.

-visualSummaryMinimumDistance: a cell array containing the paths of the representative images extracted through minimum distance.

-visualSummaryRandom: a cell array containing the paths of the representative images extracted choosing a random image of the event.

-visualSummaryUniform: a cell array containing the paths of the representative images extracted choosing temporally equidistant images.

9.1.2. extractCNNFeatures

This function uses the ConvolutionalNN provided by Caffe to extract features for the given set of images.

Input:

-path: string containing the path of the images set to analize.

Output:

-features: matrix containing the features vectors of the images of the specific path.

9.1.3. extractNF

This function extracts all the images names of an specific path.

Input:

-path: string containing the images path.

-features: matrix containing the CNN features of all the images of the path.

Output:

-featuresWithName: matrix containing the CNN features of all the images of the path and with the names of the images at the first column of the matrix.

9.1.4. normalizeL2

This function normalizes the input matrix with the L2 method.

Input:

-X: matrix of CNN features.

Output:

-X: normalized matrix of CNN features.

9.1.5. obtainAutomaticCluster

This function group into clusters the features of the daily images.

Inputs:

-featuresMatrix: is a matrix with all the features of all images of the daily extraction. The first column contains the name of the image.

-cutoff: is a value between [0 2] that is used to determine at what distance between images a cluster has to be considered (1.15 best experimental value).

-method: an string indicating the method to apply for the linkage between features. 'method' can be: average, centroid, complete, median, single, ward, weighted.

Output:

-clust_auto: a cell where every position contains a vector with the indices of the images that belong a cluster.

-numClusters: is a scalar that indicates the number of resultant clusters.

9.1.6. orderClustersTemporally

This function applies the division algorithm to obtain the ordered events taking into account the temporal information of the images name.

Inputs:

-nameImages: vector containing the name of the images.

-clust_auto: cell array with the automatic event clustering.

-N: scalar indicating the minimum number of images that must contain an event.

Output:

-orderedCluster: two-column matrix. The first column contains the images name, and the second one contains the event id which belong to the related image.

-boundariesAutomaticCluster: vector containing the events boundaries of the ordered cluster.

9.1.7. orderClusters

This function applies the division algorithm to obtain the ordered events.

Inputs:

-nameImages: vector containing the name of the images.

-clust_auto: cell array with the automatic event clustering.

-N: scalar indicating the minimum number of images that must contain an event.

Output:

-orderedCluster: two-column matrix. The first column contains the images name, and the second one contains the event id which belong to the related image.

dM

This function computes the distances matrix of the features vectors.

Input:

-features: is a matrix containing the CNN features of the images.

Output:

-distances: is a matrix containg the distances between all the images of the input matrix.

9.1.8. randomWalkClustering

This function extracts the representative image of an event applying the random walk algorithm.

Input:

-distances: a distance matrix of the images of the event. The first column has to be the name of the image.

Output:

-im: an string with the path of the representative image of the event. You can extract the less repetitive image of the event taking out the comments of the code.

9.1.9. minimumDistanceClustering

This function extracts the representative image of a cluster based on the distances matrix.

Input:

-distances: a distance matrix of the images of the event. The first column has to be the name of the image.

Output:

-im: an string with the path of the representative image of the event. You can extract the less repetitive image of the event taking out the comments of the code.

9.2. Evaluation functions

9.2.1. resultsClusterValidation

This function computes the Jaccard index along a rank of cutoff values.

Inputs:

-path: string containing the path of the images to summarize. Example of path: path='/imatge/rmestre/work/Images/testImages/Petia2';

-excel_filename: string containing the groundtruth file to analize. Example: excel_filename='GT_Petia2.xls';

-featuresMatrix: matrix containing the CNN features of the set images. The first column contains the name of this images.

-method: string indicating the method of linkage for the clusterdata function.

-orderFlag: scalar. If it is 0, the division algorithm will not be executed. If its value is 1, the temporally division will be applied, and if its value is 2 the division algorithm will be applied without taking into account the temporal information.

-normFlag: boolean. If it is true, the normalization will be applied.

-N: scalar that indicates the minimum number of images that must contain an event.

Output:

-cutoff: vector of cutoff values.

-numClusters: for the cutoff values, the related number of clusters that every value cause.

-JI: for the cutoff values, the related Jaccard index.

9.2.2. analizarExcel_Narrative

This function extract the groundtruth events of an excel file.

Inputs:

-excelfile: string containing the name of the excel file.

-path: string indicating the path of the set of images.

-N: scalar indicating the minimum number of images that must contain an event. If there is an event annotated with less than N images, the code will remove this images.

Outputs:

-events: cell array containing the groundtruth events.

-clustersId: vector indicating in which cluster belongs an image.

-cl_limGT: vector with the boundaries of every event.

-sum: number of images in the set.

-imagesRemoved: vector with the removed images name.

9.2.3. JaccardIndex

This function extracts the Jaccard index between the groundtruth and the automatic segmentations.

Inputs:

-clust_man_ImagName: cell array with the groundtruth segmentation.

-clust_auto_ImagName: cell array with the automatic segmentation.

Outputs:

-JImean: Jaccard index mean.

9.3. <u>Other functions</u>

9.3.1. resizeImages

This function resizes all images in a specific path to 490x472 pixels.

Input:

-path: is an string containing the path of the images to resize. Example: path='C:\Users\RICARD\Universitat\TFG\SetsParaRicard\petia_2';