Spatio-Temporal Wardrobe Generation of Actors' Clothing in Video Content

Florian Vandecasteele[1], Jeroen Vervaeke[1], Baptist Vandersmissen[1], Michel De Wachter[2], and Steven Verstockt[1(✉)]

[1] ELIS Department - Data Science Lab, Ghent University IMinds, Sint-Pietersnieuwstraat, 41, 9000 Ghent, Belgium
{florian.vandecasteele,steven.verstockt}@ugent.be
[2] Appinness, Hertshage 10, 9300 Aalst, Belgium

Abstract. In this paper, we propose a methodology for spatio-temporal wardrobe generation for video content. The main goal is to suggest relevant matches between clothes worn by actors and images originating from a set of e-commerce clothing sites. The semi-automatic generation of fine-grained spatial metadata for each video sequence is based on shot detection, keyframe detection, feature matching and clothing type classification based filtering. The result of this annotation process is a spatio-temporal database consisting of videos and the corresponding actor clothing. This database can be queried in various ways depending on the intended target application.

Keywords: Video summarization · Shot detection · Clothing annotation · Metadata enrichment · Deep learning

1 Introduction

The clothing industry amounts for one of the most important selling segments in e-commerce worldwide[1]. Furthermore, the majority of online shopping happens spontaneous, often based on popular trends displayed on television (TV) [1], i.e., consumers actively search based on what their role models wear on TV. Currently, no automatic tool exists that is able to link the digital wardrobe of an actor to the same clothes visually displayed on a TV screen. The consumer's search for similar clothing is mostly based on the actors branding and a rough textual description of the type and color. Such strategy does not enable people to efficiently find similar clothes. In this paper, we propose a solution for spatio-temporal recognition and annotation of clothing in video content. Our system significantly reduces the required efforts for consumers to find interesting items. The main focus of this paper is on the spatio-temporal recognition of clothes in video shots. In addition, we also introduce a novel keyframe selection mechanism, effectively reducing the set of relevant frames. By limiting the amount of

[1] http://bit.ly/1F2zC9M.

M. Kurosu (Ed.): HCI 2016, Part III, LNCS 9733, pp. 448–459, 2016.
DOI: 10.1007/978-3-319-39513-5_42

keyframes, while still maintaining all relevant information present in the video, significant time, needed to label and predict clothing matches for a video shot, can be gained.

The remainder of this paper is organized as follows. Section 2 gives an overview of the proposed framework. Subsequently, Sect. 3 proposes the workflow of the video summarization algorithm. Furthermore, Sect. 4 focuses on the clothing annotation and recognition. In Sect. 5, we present the tool for manual verification of the clothing tagging and we demonstrate our approach on an exemplary TV application. Finally, Sect. 6 lists the conclusions.

2 Framework

The proposed architecture, shown in Fig. 1, consists of 3 main building blocks: (a) a low-complexity video summarization algorithm; (b) a region-of interest (ROI) selection based on face detection; and (c) feature-based clothing object annotation. The generated results are stored in an XML or JSON file depicting the different clothing matches, the video timestamp (frame number) and the spatial location in the keyframe.

The video summarization, i.e., the first building block, is subdivided into three steps, namely: a keyframe selection mechanism (Sect. 3.1), followed by a quality analysis in order to find the best representative frame of a shot (Sect. 3.2), and finally a keyframe similarity mechanism (Sect. 3.3) to reduce the amount of similar keyframes. The final set of distinct keyframes is used as input for the second part of our architecture, i.e., the face detection module (Sect. 4.1), and clothes region selector (Sect. 4.2). The predicted ROI (depicting the region of clothes) and the information concerning the coordinates of the face are used as inputs for the final step of our system, namely the clothing matching based on a coarse to fine strategy (Sect. 4.4).

3 Video Summarization

The automatic understanding and summarization of video content is a challenging problem. The main goal is to reduce the amount of data (frames) by filtering out redundant and unnecessary frames, while preserving only those frames, distinctive and essential to capture the entire video content. A lot of research has been done in the area of video summarization. Ajmal et al. [2], for example, give an overview of the different techniques and classification methods that are commonly used in literature, e.g., feature classification [3], clustering [4], shot selection [5] and trajectory analysis [6]. The focus of this paper will be on shot selection and detection [7–9]. The main research challenge is to properly cope with camera movements within a shot. In order to tackle this issue, we present a grid-based histogram shot detection method. In combination with an additional quality analysis step, this technique gives us the best frame for each shot, i.e., the frame with the most optimal lightning and contrast. Finally, we generate a list of keyframes, where each frame represents a new viewpoint of a scene. A camera

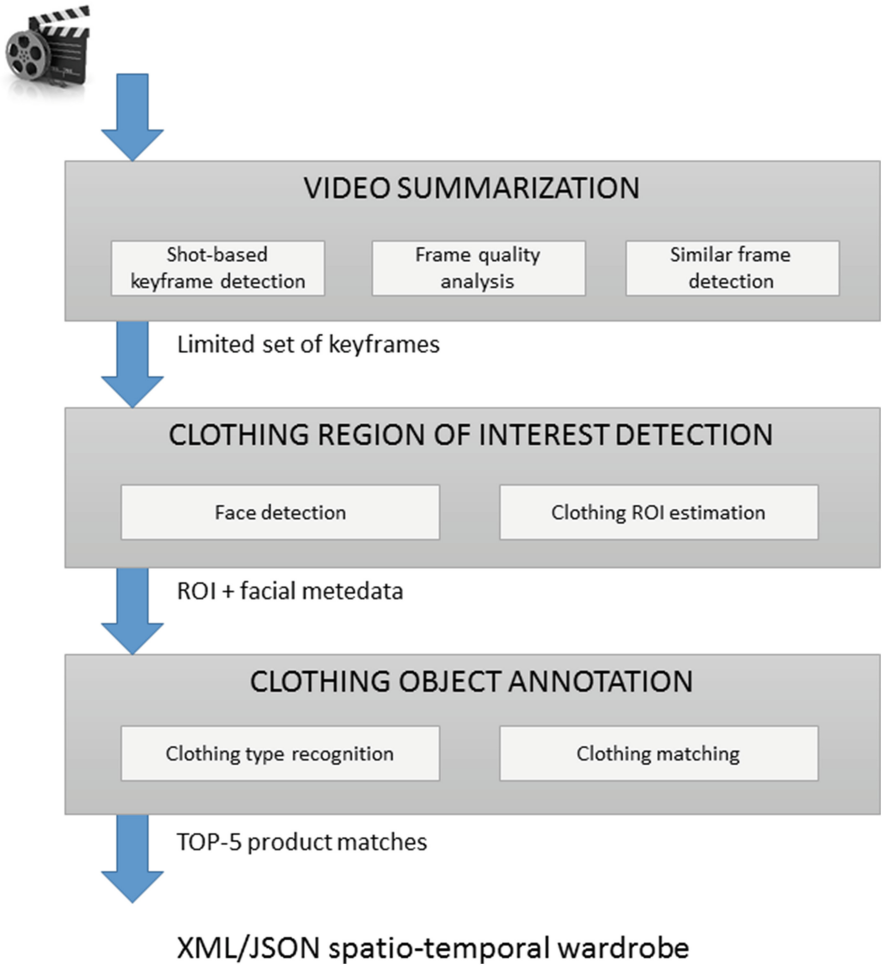

Fig. 1. Modular architecture for spatio-temporal wardrobe generation of video content

shift between one actor to an other, for example, results in two shots. The proposed mechanism is suitable for non-fixed camera movements and works without significant content knowledge. Compared to Baraldi et al. [10], which recently proposed a novel hierarchical clustering based video summarization algorithm, we focus on finding the best representative keyframe within each shot, while they focus on segmenting broadcast videos into coherent shots/scenes.

3.1 Keyframe Selection

The proposed approach for keyframe selection is based on local histogram analysis on a 5-by-5 grid. Compared to state-of-the-art temporal shot selection algorithms [7–9] it is able to cope with fast camera movements, zoom gestures,

gradual shot transitions, and similar scene discrimination. An evaluation on a dataset of 400 scenes from different kinds of commercial video content (cooking programs, TV series, clothing styling programs, film trailers) results in a recall of 96 % and a precision of 90 %. The proposed algorithm performs especially well on detecting hard and gradual shots, making it suitable to process multiple types of video content. The main limitation of the histogram grid based shot detection is that it fails to cope with similar scenes. However, this problem is tackled in Sect. 3.3. Furthermore, in order to cope with gradual shots, such as blends and fades, the amount of frames after the last detected transition is counted. By comparing this amount to an experimentally defined threshold, we decide whether to consider it as a new or the same transition. In this way, we successfully manage to detect both gradual and abrupt scene transitions.

3.2 Keyframe Quality Analysis

The clothing recognition algorithm only uses the most representative keyframe per shot as input, i.e., the frame predicted to have the highest quality within each shot. Therefore, a weighted combination of three no-reference quality metrics (blur, contrast and brightness) is computed and used as a real-time image quality measure. Currently, multiple image quality metrics are available [11]. However, most of the existing metrics require a reference image or are not suitable for real-time quality measurement. The proposed no-reference blur, contrast and brightness metrics are evaluated on a variety of television programs. The blur-metric is based on an edge strength analysis. The sharpness of an image is computed by summing the partial differentiate in vertical and horizontal direction. Finally, the shift and spread of the histogram are analyzed to evaluate the brightness and contrast of the image. The combination of those metrics achieves performances that match the human perception for best frame selection, which is proven by a subjective evaluation on our dataset of TV shots.

3.3 Similar Frame Detection

To ensure a robust, reliable, and scalable setup, and in order to avoid recurrence of similar queries, i.e., processing overhead, the detected keyframes are filtered on similarity. We note that in this context similar or near duplicate frames are defined as frames originating from the same scene but showing a different viewpoint or change in illumination.

Numerous approaches detecting near-duplicate images or frames have been published in literature. Chum et al. [12], for example, propose a color histogram local sensitivity hashing (CH-LSH) as the best method for near duplicate image search. They assume that the Euclidean distance between two images (described by their color feature vectors) is a meaningful measure of image similarity. However, the Euclidean distance is found not to be robust and local occlusions can cause a significant change in an image's color histogram. Tasdemir et al. [13] describe a motion vector based approach for copy detection in video content. Experimental results, however, do not seem promising in our context. Sarkar et al. [14] compare

YCbCr histograms, Compact Fourier-Mellin Transform (CFMT) and Scale Invariant Feature Transform (SIFT) for video fingerprinting and large scale similar video retrieval. Their results show that CFMT features perform best for providing quick, accurate retrieval of duplicate videos. Our proposed near duplicate recognition system follows a similar strategy and is built upon LIRE, a light weight open-source library for content based image retrieval (CBIR) [15].

The best results for similar keyframe comparison are achieved with a combination of color and edge directivity descriptor (CEDD) [16] and the fuzzy color and texture histogram feature (FCTH) [17]. The combined compact descriptors mentioned by Kumar et al. [18] show that the fusion of several smaller local descriptors improves the overall results for image indexing and retrieval. Our combination of CEDD and FCTH features for duplicate detection is evaluated on several video sequences, starting from the generated keyframes as described in Sect. 3.1. Finally, we end up with an automatically selected representative set of keyframes. This set could be visualized in an application, effectively summarizing any video. Those keyframes are then used as input for the clothing recognition mechanism, which is proposed in Sect. 4.

4 Clothing Recognition and Annotation

In recent years, the research community has actively been focusing on the automatic classification and detection of clothes in digital content [19–21]. Unfortunately, due to the large visual differences between images on e-commerce websites (photographed on a clean, white background with clear lighting conditions) and those retrieved from in-the-wild videos, this problem is still largely unsolved. In this paper, we propose a novel algorithm to find the best match between clothes in a keyframe and related images found on e-commerce websites. The proposed methodology uses a coarse to fine strategy, which is inspired by the current search mechanisms of popular e-commerce sites. By adding metadata elements like gender, color, texture, and clothing type we filter clothing indexes and predict the most likely matches.

4.1 Face Detection

On each keyframe, returned by the video summarization tool, a face detection algorithm is executed. Face detection and face metadata generation is still an active research domain. However, nowadays multiple commercial solutions are available (e.g. Betaface [22] and openbiometrics [23]). Since improving face detection algorithms is out of scope for this paper, we make use of the commercial Face++ algorithm [24]. An evaluation is done by first comparing the output of several commercial detectors, and additionally evaluating the estimated face coordinates. Furthermore, Face++ outperforms the other approaches on age and gender classification. Both features are used as filters on our indexed clothing datasets to guide and improve the clothing search process. Important to mention is that further improvements on face descriptors would be beneficial to optimize our global architecture.

4.2 Clothing Object Segmentation

Due to different body poses, clothing colors and textures, clothing object segmentation is still an active research problem. Simo-Serra et al. [25], for example, propose a promising CRF model to parse clothes in an image. However, similar to the majority of state-of-the-art approaches, this method fails when the frame only contains upper body parts. For this reason, we propose an alternative approach which is able to cope with this issue.

Based on the position of the detected face we construct a region-of-interest (ROI) that can be used for the proper selection of a clothing patch. The automatic ROI selection takes into account the Face++ results for position, orientation and dimension of the face to proportionally estimate the most probable clothing location in the image. The coordinates of the ROI are based on the position of the lowest neck pixels. By computing a histogram of the face, we obtain a descriptor of skin like pixel colors. These values can be compared to those of the neck pixels, resulting in an estimation for accurate clothing boundaries.

Given a ROI depicting the clothing object(s), our architecture contains three different methods to perform clothing patch generation. Our first method makes use of a superpixel segmentation of the ROI. These superpixels are groups of perceptually meaningful pixel regions which reduce the image complexity into homogeneous regions that align well with object boundaries. Based on a thorough evaluation of different superpixel strategies, simple linear iterative clustering (SLIC) [26] is currently found to perform best. Finally, we merge the superpixels into our clothing query object using a graph based approach [27], focusing on texture, Lab color and location similarities. The main problem with this automatic approach, however, is its inability to cope with differing colors and textures withing a same clothing piece. The merging of superpixels can potentially result in a patch without any meaningful part of the clothing. To avoid this problem, a second approach is developed based on taking a proportional crop of the given ROI. The third and final approach relies on manual input selection by a domain expert. Future work will focus on further optimizing/automizing the clothing segmentation.

4.3 Clothing Type Recognition

Even though color and texture information derived from a clothing patch are essential pieces of information to detect the exact clothing wear, determining the type of clothing can also greatly aid to reduce the target search space. In this subsection we will briefly describe our approach towards automatic clothing type recognition. Important to mention is that we use one of the pretrained Visual Geometry Group (VGG) networks of Simonyan et al. [28], that has proven to be suitable for general image understanding and different classification tasks. The VGG network is a successor of Alexnet, significantly increasing the depth of the convolutional network (up to 19 layers). While Alexnet achieved a top-5 accuracy of 80 % on the ImageNet test set, VGG improved this to 89 % by using a deeper network with very small filters (3x3).

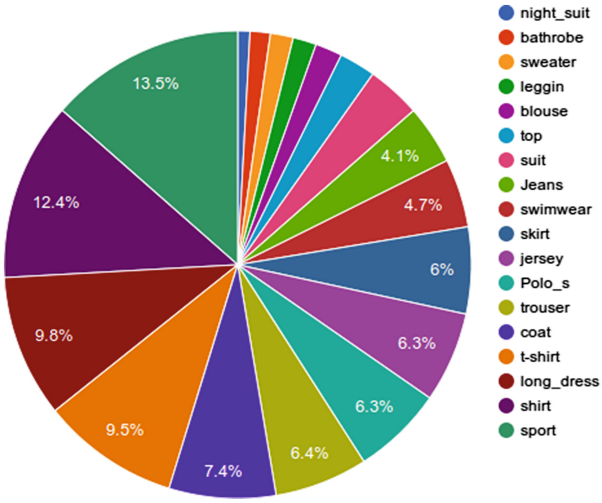

Fig. 2. Class distribution of collected clothing dataset

The dataset was collected from Zalando, containing 7210 samples, distributed over 18 different classes. Figure 2 shows the class distribution. We replace the 1000-dimensional softmax layer of the 16-layer pretrained VGG network with an 18-dimensional softmax layer to adapt the network to our needs. First, 10 % of the dataset is randomly left out and used as a validation set, while the other 90 % represent the training set. Images are preprocessed by subtracting the mean image values and generating randomly cropped, mirrored and shifted images to augment the dataset. Minibatch gradient descent in combination with momentum and a degrading learning rate is used to fine tune the network. This network was tested on keyframes originating from different television shows. The clothing type classification using an entire image as input achieves a top-1 accuracy of 35 % and a top-3 accuracy of 47 %. Results on the region of interest based classification task results in a top-1 accuracy of 15 % and a top-3 accuracy of 49 %. The accuracy of the clothing type classification task is relatively low, but similar results are obtained in state-of-the-art approaches [19,20]. This accuracy shows the necessity for manual verification of the classification results. This will be further explained in Sect. 5. In future work, we will fine-tune our network for the classification of the clothing type by incorporating the manual verification data in a feedback loop.

4.4 Clothing Matching

Finally, the algorithms described in Sects. 4.1 to 4.3 result in several input parameters for the coarse to fine clothing matching. First of all, we incorporate the Face++ metadata as an index filter, i.e., we use the returned attributes to filter the clothing indexes on gender and age. By incorporating this information, we

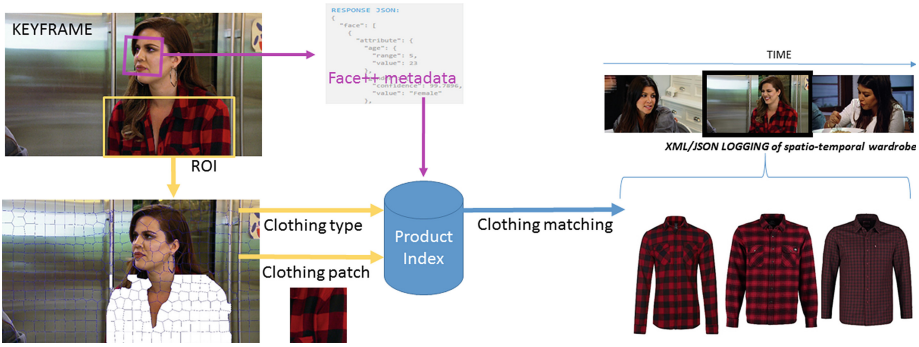

Fig. 3. Overview of the proposed clothing matching strategy - experimental results based on "Keeping Up with the Kardashians".

can already significantly improve the results. Furthermore, we also generated LIRE based clothing indexes for each predefined clothing type (similar to those of the classification task). Based on the recognized clothing type, we use global image features of the clothing patch to find the best visual match within the particular index. Finally, by reranking the CEDD [16], FCTH [17] and PHOG [29] features, we are able to generate decent results, i.e., the majority of top-3 matches in our test cases are subjectively marked as relevant by our lead users. Figure 3 gives a general overview of all different steps involved in finding a clothing match.

5 Demonstrator

In order to facilitate manual verification of the clothing type and the proposed clothing patch, we have built a validation tool (shown in Fig. 4). First, gender and clothing type is predicted based on an input frame. Wrong predictions can

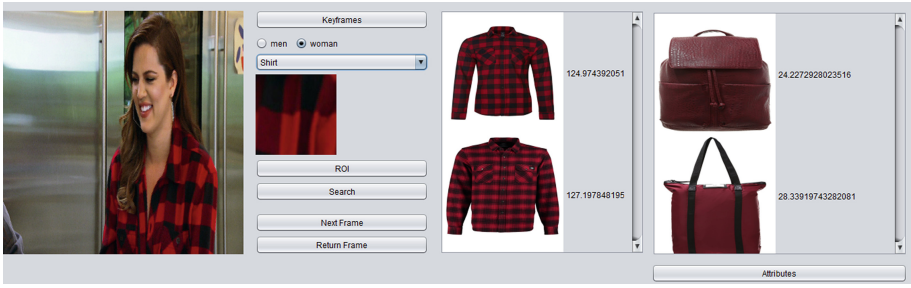

Fig. 4. Tool for manual verification and correction of the clothing match showing the keyframe (left), the best clothing matches based on the adaptable gender and clothing type estimation (middle), and a proposal of similar fashion accessories (right).

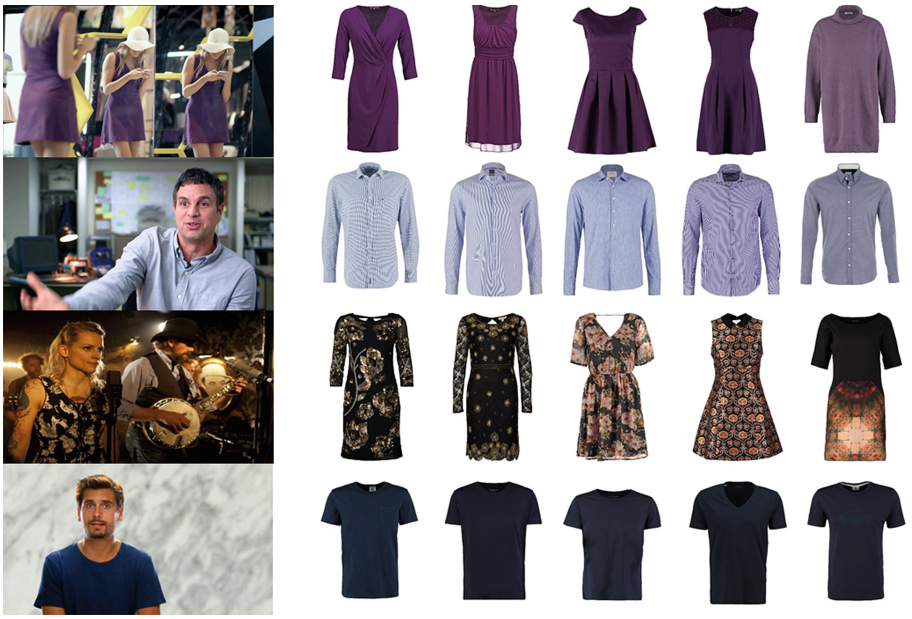

Fig. 5. Corresponding keyframes (left) and top-5 matches (right) for actors wardrobe based on the Zalando dataset.

easily be corrected by the evaluator. Second, a clothing patch is suggested. If the proposed patch is incorrect, a new region of interest can be drawn manually. Third, the search query shows the best matches with their corresponding matching rate. The lower the rate, the higher the visual features correspond. Furthermore, some fashion accessories are proposed that are similar to the color and texture of the best clothing match. This will facilitate the e-commerce shopping by enabling shopping for a complete look. Finally, The matches from our clothing tagging are stored in XML/JSON format and they are labeled with the corresponding video timestamp, location, matching accuracy, and the actors id. Actor-based querying can be performed using the trainable Face++ recognition.

The XML/JSON spatio-temporal actor wardrobes can be used in a wide range of applications, such as second screen TV shopping apps and video clothing search engines. To show the effectiveness of our clothing tagging framework, we show some results of our semi-automated tool in Fig. 5. There is no straight forward way to evaluate the clothing matches because there is no exact match in the clothing dataset. However, subjective evaluation with our lead users have shown that the given matches are highly relevant.

6 Conclusion and Future Work

In this paper we proposed a novel methodology for linking clothing of actors to their corresponding keyframe. The proposed spatio-temporal actor wardrobes

help improving the viewing experience and facilitate e-commerce shopping by making the TV content interactive and allowing you to shop what you see.

Currently, it is not possible to fully automate the clothing recognition pipeline. However, based on the proposed methodology a large reduction of the manual tagging effort is already achieved. Future work will optimize the clothing type classification and the overall scene understanding. Furthermore, a larger evaluation of the proposed shot detection and keyframe generator will be performed. Finally, it is important to stress that the continuous evolution in fine grained image understanding and classification is expected to further improve the proposed methodology in the upcoming years.

Acknowledgments. SpotShop (http://www.iminds.be/nl/projecten/2015/10/01/spotshop) is a research project facilitated by iMinds and funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT).

References

1. Liaukonyte, J., Teixeira, T., Wilbur, K.C.: Television advertising and online shopping. Mark. Sci. **34**(3), 311–330 (2015)
2. Ajmal, M., Ashraf, M.H., Shakir, M., Abbas, Y., Shah, F.A.: Video summarization: techniques and classification. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (eds.) ICCVG 2012. LNCS, vol. 7594, pp. 1–13. Springer, Heidelberg (2012)
3. Wang, F., Ngo, C.-W.: Summarizing rushes videos by motion, object, and event understanding. IEEE Trans. Multimedia **14**(1), 76–87 (2012)
4. dos Santos Belo, L., Caetano, C.A., do Patrocínio, Z.K.G., Guimarães, S.J.F.: Summarizing video sequence using a graph-based hierarchical approach. Neurocomputing **173**, 1001–1016 (2016)
5. Uchihachi, S., Foote, J.T., Wilcox, L.: Automatic video summarization using a measure of shot importance and a frame-packing method. US Patent 6,535,639 (2003)
6. Qiu, X., Jiang, S., Liu, H., Huang, Q., Cao, L.: Spatial-temporal attention analysis for home video. In: IEEE International Conference on Multimedia and Expo, pp. 1517–1520. IEEE (2008)
7. Chalamala, S.R., Kakkirala, K., Dhillon, J.: A robust video synchronization method based on hierarchical shot detection. In: International Conference on Audio, Language and Image Processing (ICALIP), pp. 206–210. IEEE (2014)
8. Liu, T.-R., Chan, S.-C.: Automatic shot boundary detection algorithm using structure-aware histogram metric. In: 19th International Conference on Digital Signal Processing (DSP), pp. 541–546. IEEE (2014)
9. Thomas, S.S., Gupta, S., Venkatesh, K.S.: An energy minimization approach for automatic video shot and scene boundary detection. In: Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), pp. 297–300. IEEE (2014)
10. Baraldi, L., Grana, C., Cucchiara, R.: Shot and scene detection via hierarchical clustering for re-using broadcast video. In: Azzopardi, G., Petkov, N., Yamagiwa, S. (eds.) CAIP 2015. LNCS, vol. 9256, pp. 801–811. Springer, Heidelberg (2015). doi:10.1007/978-3-319-23192-1_67

11. Joy, K.R., Sarma, E.G.: Recent developments in image quality assessment algorithms: a review. J. Theoret. Appl. Inf. Technol. **65**(1) (2014)
12. Chum, O., Philbin, J., Isard, M., Zisserman, A.: Scalable near identical image and shot detection. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, pp. 549–556. ACM (2007)
13. Taşdemir, K., Cetin, A.E.: Content-based video copy detection based on motion vectors estimated using a lower frame rate. Signal, Image Video Process. **8**(6), 1049–1057 (2014)
14. Sarkar, A., Ghosh, P., Moxley, E., Manjunath, B.S.: Video fingerprinting: features for duplicate and similar videodetection and query-based video retrieval. In: Electronic Imaging 2008, pp. 68200E–68200E. InternationalSociety for Optics and Photonics (2008)
15. Lux, M.: Lire: open source image retrieval in java. In Proceedings of the 21st ACM International Conference on Multimedia, pp. 843–846. ACM (2013)
16. Chatzichristofis, S.A., Boutalis, Y.S.: CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 312–322. Springer, Heidelberg (2008)
17. Chatzichristofis, S., Boutalis, Y.S., et al.: Fcth: fuzzy color and texture histogram-a low level feature foraccurate image retrieval. In: Ninth International Workshop on Image Analysis for MultimediaInteractive Services, WIAMIS, pp. 191–196. IEEE (2008)
18. Praveen Kumar, P., Aparna, D., Venkata Rao, K.: Compact descriptors for accurate image indexing and retrieval: fcthand cedd. In: International Journal of Engineering Research and Technology (2012)
19. Wang, H., Zhou, Z., Xiao, C., Zhang, L.: Content based image search for clothing recommendations in e-commerce. In: Baughman, A.K., Gao, J., Pan, J.-Y., Petrushin, V.A. (eds.) Multimedia Data Mining and Analytics, pp. 253–267. Springer, Heidelberg (2015)
20. Kiapour, M.H., Han, X., Lazebnik, S., Berg, A.C., LBerg, T.: Where to buy it: matching street clothing photos in online shops. In: Proceedings of the IEEE International Conference on ComputerVision, pp. 3343–3351. IEEE (2015)
21. Nogueira, K., Veloso, A.A., dosSantos, J.A.: Pointwise and pairwise clothing annotation: combining features fromsocial media. Multimedia Tools Appl. **75**, 4083–4113 (2015)
22. Šaloun, P., Stonawski, J., Zelinka, I.: Automated face comparison with facebook friend's faces and flickr photos. In: Zelinka, I., Duy, V.H., Cha, J. (eds.) AETA 2013. LNEE, vol. 282, pp. 349–362. Springer, Heidelberg (2014)
23. Klontz, J.C., Klare, B.F., Klum, S., Jain, A.K., Burge, M.J.: Open source biometric recognition. In: IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8. IEEE (2013)
24. Fan, H., Yang, M., Cao, Z., Jiang, Y., Yin, Q.: Learning compact face representation: packing a face into an int32. In: Proceedings of the ACM International Conference onMultimedia, pp. 933–936. ACM (2014)
25. Simo-Serra, E., Fidler, S., Moreno-Noguer, F., Urtasun, R.: A high performance CRF model for clothes parsing. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9005, pp. 64–81. Springer, Heidelberg (2015)
26. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 2274–2282 (2012)

27. Yang, J., Gan, Z., Li, K., Hou, C.: Graph-based segmentation for rgb-d data using 3-d geometry enhanced superpixels. IEEE Trans. Cybern., 927–940 (2015)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv:1409.1556
29. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, pp. 401–408. ACM (2007)