GESTURE-BASED VIDEO SUMMARIZATION

D. I. Kosmopoulos[1], A. Doulamis[2], N. Doulamis[3]
[1]National Centre for Scientific Research Demokritos, Institute of Informatics & Telecommunications
[2,3]National Technical University of Athens, Dept. of Electrical and Computer Engineering
[1]dkosmo@iit.demokritos.gr, {[2]adoulam, [3]ndoulam}@cs.ntua.gr

ABSTRACT

A novel method for summarizing videos of gestures is presented. The gestures performed by the hands and the head are extracted through skin color segmentation and represented through Zernike moments. The gesture energy is calculated using the norms of the Zernike moments and monitored through time for local minima and maxima that indicate distinctive visual events and thus key-frames. The proposed scheme is not threshold-dependent and therefore the number of extracted key-frames varies according to the complexity of gesture energy variation. The applicability of the method is verified experimentally in sign language videos.

1 INTRODUCTION

Traditionally, video is represented as a sequence of consecutive frames, each of which corresponds to a constant time interval [1]. This linear representation, though adequate for playing a video in a movie, is not appropriate for the new emerging multimedia services that require new tools and mechanisms for interactive navigating video information over various networks and devices of limited bandwidth. Currently, the only way to exchange visual data among different databases is to perform video streaming or video file downloading, methods that are tedious and time consuming.

To overcome these issues, a new non-linear organization of video files should be performed exploiting media content information. Video summarization, which aims at extracting a small but meaningful abstract of a video file, can be seen as an interesting case of non-linear video description. Some first approaches for video summarization dealt with the construction of image maps and image mosaics [2], [3]. These approaches however provide satisfactory results only for scenes of simple content without complicated camera effects. In [4], a method for analyzing video and building a pictorial summary has been presented, while in [5] a fuzzy visual content representation has been proposed with application to video summarization and content based indexing and retrieval. Color and depth information have been appropriately combined in [6] to summarize stereoscopic video sequences. In [7], the class separation measure has been used to investigate the effect of the number of classes on the visual content, while in [8] an interpolation-based method for video abstraction has been proposed. Video summarization of sports events (soccer, baseball) has been reported in [9]-[11]. In [12], a novel video summarization method has been presented using perceived motion energy, while in [13] the MPEG-7 tools are used to perform the summarization task.

The above mentioned works are oriented for generic video shots of any type or for video shots of particular content, e.g., sport events. In this paper, a novel summarization approach is proposed for videos of gestures that are used for applications such as sign language communication, human-computer interaction, or even robot guidance over networks. In these types of videos, the semantic content is mainly defined as the gesture variation, i.e., the movement of hands and head, the relative location of hands and head as well as the hand-shape. As a result, direct application of video summarization algorithms to such video content cannot provide satisfactory results, i.e., close to the human perception of the scene semantics. Thus, the goal of the presented approach is to dramatically reduce the number of frames of such gestured-based videos without however losing important information as far as the meaning of the gesture is concerned. For example, in sign language communication, the goal would be that one is able to understand the semantic meaning of the "words" (i.e., gestural signs) with a very limited amount of information. This is the most important contribution of this paper. Such a summarization approach is very important for adapting transmission and delivery of these types of videos over low-bandwidth networks or terminal devices of different characteristics, such as PC's, notebooks, PDA's and mobile phones.

The proposed architecture consists of the gesture segmentation - representation module and the summarization module. Gesture segmentation is performed in our case taking into account skin color information. Then, the segmented regions are represented using the Zernike moments [14]. The adoption of Zernike moments, instead of other types of moments, such as Cartesian or Hu moments, is due to the fact that they are orthogonal and thus they have stronger representation capabilities. Due to orthogonality, the sum of the squared coefficients of the Zernike moments expresses the "energy" of the gesture shape [14]. They are also more noise resilient compared to other orthogonal moments types [15]. The pseudo – Zernike moments have been reported to perform better in the presence of noise [16] but their complex calculation introduces unwanted overhead.

The next step is to apply a key-frame extraction algorithm. A fast and effective algorithm for non-sequential video content representation is applied, which exploits the fluctuation of the gesture shape (energy) over time. In particular, the local minima/maxima of the second derivative of the gesture energy are considered as the most relevant for key-frame extraction. The proposed scheme is not threshold dependent and thus the number of extracted key-frames varies according to the complexity of gesture energy variation. Furthermore, the computational complexity of the algorithm is low and allows real-time application. Other advantage of the scheme is its ability to detect periodical gesture movements in contrast to mainstream methods, such as the [5], [6] and [8]. This is very important in gesture-based video content since usually periodic movements are presented to indicate a particular concept (e.g., in sign language).

2 GESTURE SEGMENTATION - REPRESENTATION

As mentioned above, the goal of this paper is to extract a limited amount of video frames from content, in which the gesture

motion is salient. As gesture motion, we define the movement of the head and both hands. Thus, the first step in our approach is the automatic localization of the hands and face since the hand-face relative location is also important in defining the semantic meanings of such kind of videos.

To achieve this, we firstly locate the face in the image, we train a color classifier based on the face color and then based on the modeled color we find the hand regions in the image. In this fashion we find up to three separate regions, which represent the skin of the face and both hands. These regions are represented using Zernike moments, which will be used for defining the gesture-based visual events as will be analyzed in the next section.

We assume that the depicted person (target) faces the camera with his/her upper body part captured in the image. The hands may disappear from the image, they may be occluded by each other and they may occlude the head, but we assume that the target is dressed.

2.1 *Initialization*

During initialization we seek to detect the target's face within the image, in order to (a) define the gestures associated with the head and (b) to train a skin color classifier. Elaboration in face detection methods is not within the scope of this research and therefore no theoretical issues will be addressed here.

After the detection of the target's face, using a readily available vision library, we use a part of the face region for skin color modeling. For solving this target-specific and thus reduced skin color modeling problem, we assume a single multivariate Gaussian model. The color probability density function in the Hue, Saturation and Intensity color space (selected due to similarity to human perception) is given by:

$$p(\mathbf{c} \mid skin) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{c}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{c}-\boldsymbol{\mu})^{\mathrm{T}}} \quad (1)$$

where $\mathbf{c}$ refers to the color vector and $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ are the mean vector and the covariance matrix respectively. The skin regions are obtained by thresholding the respective probability values. Depending on the application context the I (and sometimes also the S) channel may be excluded from the color probability calculation.

2.2 *Activity state representation by Zernike moments*

Using the color model extracted during initialization phase, we are able to segment the image into skin and non-skin regions resulting in a binary mask. Possible noise of the mask can be removed by applying a morphological filter (e.g., the opening operator). Then the resulting regions are processed with connected components to eliminate small false positive skin areas, keeping a maximum of three areas corresponding to the head and the two hands.

The final mask is applied to the Intensity channel of the image to obtain a masked gray-level image including only the head and the two hands. The gray-level image provides a richer representation of the current gesture than the binary masks; the employment of the latter can lead to loss of information especially in the case of different gestures with similar silhouettes, e.g., front and back image of the hand. We use the Zernike moments to represent the activity state as it is expressed by the relative position and shape of the head and the two hands. The Zernike moments have some very positive attributes that make them proper for our representation purposes and namely their noise resiliency, the reduced information redundancy and their reconstruction capability.

The complex Zernike moments of order p are defined as [14]:

$$\mathrm{A}_{pq} = \frac{p+1}{\pi} \int_0^1 \int_{-\pi}^{\pi} R_{pq}(r) \cdot e^{-jq\theta} f(r,\theta) \cdot r \cdot dr d\theta \quad (2a)$$

$$r = \sqrt{x^2 + y^2}\ ,\ \theta = \tan^{-1}(y/x),\ -1<x,y<1 \quad (2b)$$

$$R_{pq}(r) = \sum_{s=0}^{\frac{p-q}{2}} (-1)^s \frac{(p-s)!}{s!(\frac{p+q}{2}-s)!(\frac{p-q}{2}-s)!} r^{p-2s} \quad (2c)$$

p-q = even and $0 \le q \le p$.

For the purposes of this work we chose to normalize the moments around the center of gravity because it is actually the relative pose of the head to the hands that signifies change of activity. On the contrary, normalization to scale is undesired due to the fact that scale variations within the same shot signify interesting gestural events.

3 VIDEO SUMMARIZATION

The sum of the squared coefficients of the Zernike moments is used as a signature of gesture in the presented approach [14]. The sum is defined as

$$J = \sum_{p=0}^{Q} \sum_{\substack{q \le p, \\ \frac{p-q}{2} even}} \left\| A_{pq} \right\|^2 \quad (3)$$

where Q is the selected order of the moments and $\|\cdot\|$ the L_2 norm.

Then, the energy is plotted for each frame of the shot forming a trajectory, which expresses the temporal variation of the energy shape through time. Thus, selection of the most representative frames within a shot is equivalent to selection of appropriate curve points, able to represent the corresponding trajectory.

In our case, the second derivative of the shape energy for all frames within a shot with respect to time is used as a curvature measure. Local maxima correspond to time instances of peak variation of the object shape. In addition, local minima indicate low variation of the object shape.

Let us also denote as $J(k)$ the energy of shape coefficients to the kth frame of the examined shot. In this notation, we have added index k to indicate the energy dependency on frame index. Initially, the first derivative of signal $J(k)$, say $J'(k)$, is evaluated with respect to time index k. Since, however, variable k takes values in discrete time, the first derivative is approximated as the difference of shape energy between two successive frames $J'(k)=J(k+1)-J(k)$. However, the previous operator is rather sensitive to noise since differentiation of a signal stresses the high pass components. For this reason, a weighted average of the first derivative, say J'_w, over a window, is used to eliminate the noise influence. Particularly, the weighted first derivative is given as

$$J'_w(k) = \sum_{l=\alpha_1(k)}^{l=\beta_1(k)} w_{l-k} \left(J(l+1) - J(l) \right),\ \mathrm{k}=0,\ldots,M\text{-}2 \quad (4)$$

where $\alpha_1(k) = \max(0, k - N_w)$, and $\beta_1(k) = \min(M-2, k+N_w)$ and $2*N_w+1$ is the length of the window, centered at frame k. Variable M indicates the number of frames of the shot. It can be seen from (5) that the window length linearly reduces at shot limits. The weights w_l are defined for $l \in \{-N_w, N_w\}$; in the simple case, all weights w_l are considered equal to each other,

meaning that the derivatives of all frame feature vectors within the window interval present the same importance,

$$w_l = \frac{1}{(2N_w+1)},\ l=-N_w,\ldots,N_w \quad (5)$$

Similarly the second weighted derivative, $J''_w(k)$, for the k-th frame is defined as:

$$J''_w(k) = \sum_{l=\alpha_2(k)}^{l=\beta_2(k)} w_{l-k} J''(k) \quad (6)$$

where $J''(k) = J'(k+1) - J'(k)$, $k=0,\ldots,M-3$ (7)

and $\alpha_2(k) = \min(0, k-N_w)$, $\beta_2(k) = \min(M-3, k+N_w)$

As explained previously, the local maxima and minima of J'' are considered as appropriate curve points, i.e., as time instances for the selected key-frames. Note that J'' is a discrete time sequence. Hence, the local maxima and minima can be estimated as the union of two sets $X = X_M \cup X_m$; the X_M contains the time instances of frames corresponding to the local maxima of J'', while the X_m the time instances of local minima of J''. The sets X_M and X_m are estimated as follows

$$X_M = \{k : J''(k-1) < J''(k) \,\&\, J''(k) > J''(k+1)\}$$
$$X_m = \{k : J''(k-1) > J''(k) \,\&\, J''(k) > D(k+1)\} \quad (8)$$

4 EXPERIMENTAL RESULTS

The applicability of the method has been verified through summarizing more than 300 sign language video shots with very positive results using a library of sign gestures. For face detection we have employed the Haar face detector due to its low computational complexity and high success rate (the boosted cascade scheme speeds up the face detection, the rotated Haar-like features and the post optimization procedures increase recognition rates [13], [16], [18]). This functionality has been implemented in the OpenCV computer vision library, which we use. The sensitivity of the method to big rotations is known and therefore we have assumed head rotations smaller than 30 degrees (can be alleviated by using methods such as inserting the Haar-like facial features into deformable graphs see, e.g., [19]). For color training we have used a rectangle within the identified face region and we have employed all color channels, assuming no dramatic change in the illumination during shot acquisition. We have excluded pixels with low intensity and low saturation values due to their instability. The observed results have been of the same high quality even if we excluded the intensity channel. We have used moments of up to 17th order.

The presented scheme for gesture-based video summarization has been evaluated using video content of sign language. Figures 1(a, c, e) illustrate three typical shots of such a file representing the signs for the words “amount” and “assume”. To illustrate that our method extracts the key-frames only based on the gesture variation, another implementation of the “assume” word is made in Figure 1(e). As is observed, the duration of the second “assume” sign is longer than the first one and the pose of the target is also different.

The key-frames extracted for all the three cases are shown in Figures 1(b, d, f). It is evident that by looking only at these key frames, the entire meaning of the shot is retained. In the case of the “assume” sign the periodic movement of the hands has been captured in the key frames (see the third and fifth key frame). In addition, the same content is extracted regardless of the shot duration or the background condition since only gesture variations are taken into account. Similar results have been extracted for all

5 CONCLUSIONS

In this work a novel method for summarization of gesture videos has been presented. The key frames have been extracted using second derivative of the gesture “energy” within a time window, which has been calculated through Zernike moments coefficients. The results in sign language videos are very promising. An extension of the proposed methodology may include within its scope the summarization of image sequences or videos that depict action through gestures that are associated with more general semantics (e.g., violent scenes in movies).

6 REFERENCES

[1] B. Furht, S.W. Smoliar and H. Zhang, *Video and Image Processing in Multimedia Systems,* Kluwer Acad. Publ., 1995.

[2] M. Irani and P.Anandan,” Video indexing based on mosaic representation,” *Proc. of IEEE*, Vol. 86, No. 5, pp. 805-921, 1998.

[3] N. Vasconcelos and A. Lippman, “A spatiotemporal motion model for video summarization,” *Proc. of IEEE CVPR*, pp. 361-366, Santa Barbara, CA, June 1998.

[4] M. M. Yeung and B.-L. Yeo, “Video visualization for compact presentation and fast browsing of pictorial content,” *IEEE Trans. on CSVT*, Vol. 7, No. 5, pp. 771- 785, October 1997.

[5] A. Doulamis, N. Doulamis, and S. Kollias, "A fuzzy video content representation for video summarization and content-based retrieval," *Signal Processing*, Vol. 80, pp. 1049-1067, June 2000

[6] N. Doulamis, A. Doulamis, Y. Avrithis, K. Ntalianis, S. Kollias, "Efficient summarization of stereoscopic video sequences," *IEEE Trans. on CSVT*, Vol. 10, No. 4, pp. 501-517, 2000.

[7] A. Hanjalic and H. Zhang, "An integrated scheme for automated abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. on CSVT,* Vol. 9, No. 8, pp. 1280-1289, 1999.

[8] N. Doulamis, A. Doulamis and K. Ntalianis, “An Optimal Interpolation-based Scheme for Video Summarization,” *IEEE Inter. Conf. on Multimedia and Expo (ICME)*, Lausanne, Switzerland, August 2002.

[9] D. Tjondronegoro, Yi-Ping Phoebe Chen and Binh Pham, “Highlights for more complete sports video summarization,” *IEEE Multimedia,* Vol. 11, Oct.-Dec. 2004.

[10] A. Ekin, A.M. Tekalp, and R. Mehrotra, “Automatic soccer video analysis and summarization,” *IEEE Trans. on Image Processing,* Vol. 12 , pp. 796-807, July 2003.

[11] G. Caccia, R. Lancini, and S. Russo, “Algorithm for summarization and key extraction in athletic video,” IEEE Inter. Workshop on Multimedia Signal Processing, pp. 229-232, 2002.

[12] T. Liu, H. J. Zhang and F. Qi, “A novel video key-frame-extraction algorithm based on perceived motion energy model,” *IEEE Trans. on CSVT,* Vol. 13, pp. 1006-1013, 2003.

[13] Jae-Ho Lee, Gwang-Gook Lee, and Whoi-Yul Kim, “Automatic video summarizing tool using MPEG-7 descriptors for personal video recorder,” *IEEE Trans. on Cons. Elect.*, Vol. 49,pp. 742-749, Aug. 2003.

[14] R.Mukundan,K.R.Ramakrishnan, Moment Functions in Image Analysis: Theory and Applications, World Scientific, Singapore, 1998.

[15] C. The, R. T. Chin, On image analysis by the method of moments, *IEEE Transactions on PAMI*, Vol. 10, No 4, pp.496-513, 1988.

[16] A. B. Bhatia, E. Wolf, „On the circle polynomials of Zernike and related orthogonal sets, Proc. *Cambr. Phi. Society*, 50, pp. 40-48, 1954.

[17] P. Viola and J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," *in Conf. on Comp. Vision and Pattern Recognition, 2001.*

[18] Rainer L. and Jochen M., "An Extended Set of Haar-like Features for Rapid Object Detection," *in IEEE ICIP*, pp. 900-903, NY, 2002.

[19] Z. Yao, H. Li, “Tracking a Detected Face with Dynamic Programming”, Work. *Face Processing in Video*, WA, 2004

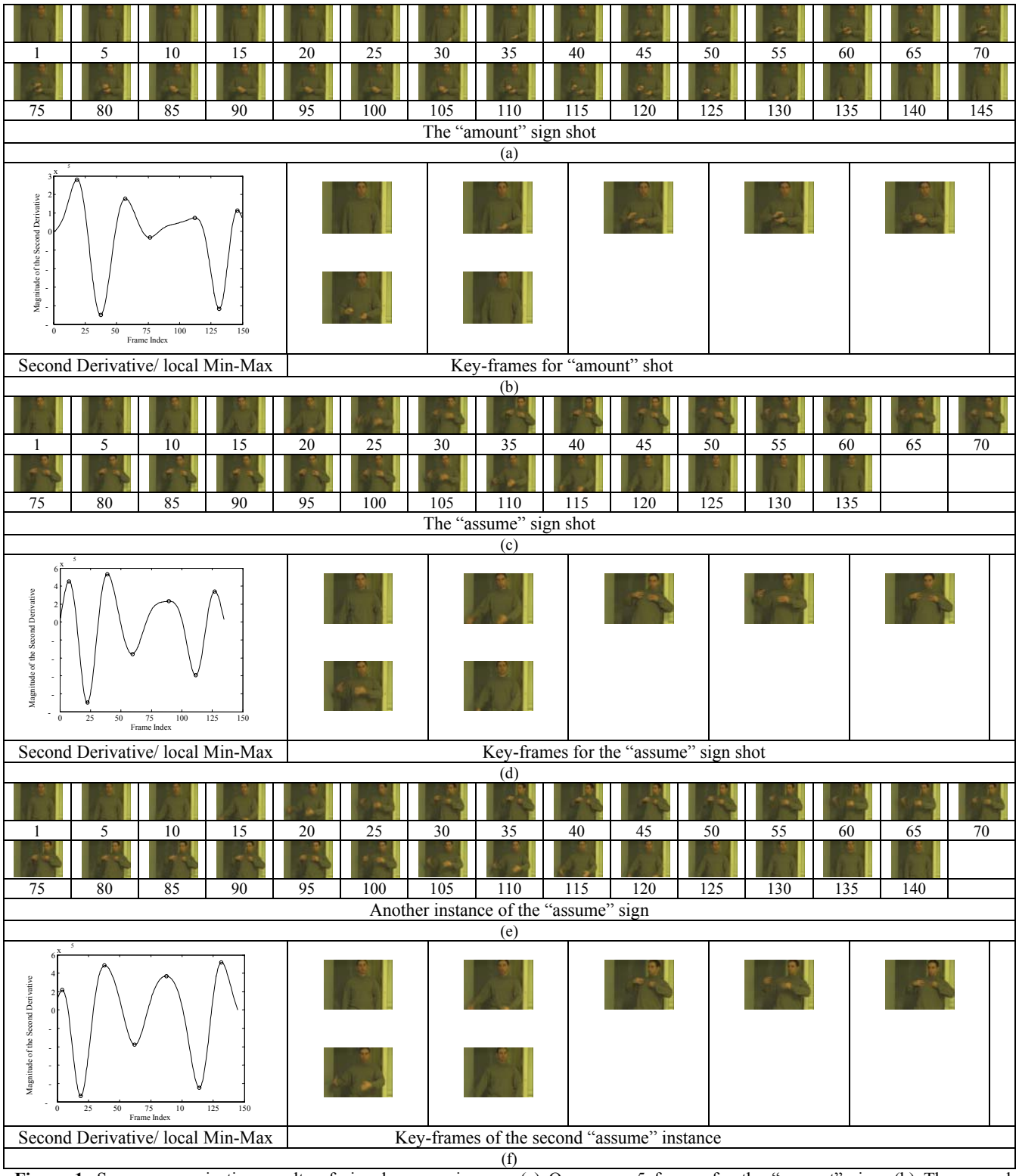

Figure 1: Some summarization results of sign language images. (a) One every 5 frames for the "amount" sign. (b) The second derivative and the local maxima/minima for the "amount" sign. (c) One every 5 frames for the "assume" sign. (d) The second derivative and the local maxima/minima for the "assume" sign. (e) One every 5 frames for another shot instance of the "assume" sign. (f) The second derivative and the local maxima/minima for the second shot instance of the "assume" sign.