VIDEO SEGMENTATION AND SEMANTICS EXTRACTION FROM THE FUSION OF MOTION AND COLOR INFORMATION

Alexia Briassouli, Vasileios Mezaris, Ioannis Kompatsiaris

Informatics and Telematics Institute
Centre for Research and Technology Hellas
Thermi-Thessaloniki, 57001, Greece

ABSTRACT

In recent years, digital multimedia technologies have evolved significantly, and are finding numerous applications, over the internet, and even over mobile networks. Thus, the video processing community has started focusing more intensively on the extraction of higher level information from multimedia data. This paper proposes a novel two-stage video processing system that aims to segment and extract semantically meaningful information, which can help achieve higher level interpretation of video. The flow fields present in the video are accumulated over several frames and their statistics are processed to derive an "activity area", that is characteristic of the type of events taking place. The color information complements the motion data, and is used for the accurate segmentation of the moving entities in each frame. The joint use of the activity area and accurate segmentation can serve as a first step to the further semantic interpretation of the video, including the recognition and accurate localization of moving objects of interest. We present experiments that demonstrate the effectiveness of our method for real videos.

Index Terms— motion analysis, semantic analysis, video signal processing, image color analysis, image segmentation

1. INTRODUCTION

The rapid development of digital multimedia and its widespread use in everyday applications can become overwhelming if this information is not analyzed and processed appropriately. Until recently, digital video had been processed only on the low-level, with features being extracted, but not interpreted. Lately, research is focusing more and more on the semantic processing of digital multimedia, which aims to extract higher level information from the available data [1], [2], [3].

Recent approaches to semantic video analysis include [4], where video semantics are extracted for the purpose of indexing. Here, the association of low-level representations and high-level semantics is formulated as a probabilistic pattern recognition problem and is addressed with the introduction of a factor graph framework. In [5], domain knowledge is combined with low-level object features and spatial descriptions, realizing an ontology-aided video analysis framework. With respect to event detection, existing approaches include among others [6], where a framework for event detection in broadcast video is developed.

In this paper, a novel method for motion estimation and segmentation is presented, which can lead to the extraction of semantic information from videos. The motion is estimated via the Lukas Kanade optical flow method [7], as it has been shown to provide reliable flow estimates between pairs of frames. These estimates are accumulated and processed (see Sec. 2), in order to extract "activity areas" (Fig. 2), which show the motion history, and thus provide a good indication of the kind of activity that is taking place in the frames that are examined. This result can already be used to extract semantics, e.g. by using it to classify temporal video segments to predefined classes, but we proceed to further analyze the color in the video, since it is also a valuable source of information that can be semantically meaningful. At each frame, we separate the possible activity area from the area that is always immobile (background). For the case of a moving camera, camera motion can be compensated for in a pre-processing stage, and our method can be applied to the resulting video. We apply mean-shift color segmentation [8] in each of these regions, to separate the color layers present in the background and the activity area. By comparing the resulting colors in the background and the activity pixels, we can decide which parts of the activity area belong to the background in each frame. This leads to the accurate segmentation of the moving objects in each frame, since their color does not match that of the background. This final result can be used to efficiently extract semantics, such as the actions taking place, and the moving entities detected in it.

This paper is organized as follows. In Sec. 2 we theoretically derive a technique for the processing of the flow estimates and the extraction of the activity areas. In Sec. 3 we briefly describe the color segmentation and comparison processes, which, along with the extracted activity areas, lead to the effective segmentation of the moving objects in each frame. In Sec. 4 we present an example of how the activity areas previously derived can be directly used for extracting some high level semantics for the video. Finally, in Sec. 5 we present experiments with real sequences, where the effectiveness and accuracy of our motion and color analysis approach is demonstrated. Conclusions and future work are discussed in Sec. 6.

2. ACTIVITY AREA EXTRACTION

Motion estimation is performed between pairs of frames in the spatial domain using the Lukas Kanade optical flow algorithm, which computes the illumination variations between pairs of frames [7], under the assumption of constant luminance. This results in flow mainly at the borders of the moving objects, which does not suffice to characterize the motion being performed, or to extract the moving object. Additionally, in practice there are always slight illumination changes in a scene, which introduce noise in the flow estimates. We present a method that actually takes advantage of the noise in the velocity estimates between pairs of frames to detect activity in the video. The general nature of the noise in the flow allows it to be satisfactorily modelled by a Gaussian distribution, particularly as it is accumulated over many frames. The velocity estimates that are

caused by actual motion deviate from the Gaussian model, since the object motion is quite different from random illumination fluctuations. This can be expressed by the following hypotheses, for the velocity estimates in frame k:

$$\begin{aligned} H_0 : v_k^0(\bar{r}) &= z_k(\bar{r}) \\ H_1 : v_k^1(\bar{r}) &= u_k(\bar{r}) + z_k(\bar{r}). \end{aligned} \quad (1)$$

Under H_0 we have a velocity estimate at pixel $\bar{r}$, in frame k that is just noise, introduced by illumination variations. Under H_1, pixel $\bar{r}$ has velocity $u_k(\bar{r})$, which is also corrupted by additive noise $z_k(\bar{r})$.

Since the noise $z_k(\bar{r})$ follows a Gaussian distribution, we can detect which $v_k(\bar{r})$ correspond to a pixel that is actually moving, by simply examining the non-gaussianity of this data [9]. The classical measure of the non-gaussianity of a random variable $\mathbf{y}$ is its kurtosis, which is defined by:

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2. \quad (2)$$

A Gaussian random variable has $E\{y^4\} = 3(E\{y^2\})^2$, so its kurtosis is equal to zero. To find the activity areas, we accumulate the noisy inter-frame velocity estimates of each pixel over the video frames, and estimate their kurtosis. In order to determine a meaningful number of frames, i.e. frames during which there is significant activity, we initially collect 10 frames. We then continue accumulating frames, comparing the new flow estimates with the mean of the previously collected flows. We consider that a "significant" event has occurred when the new flow estimate is higher than the standard deviation of the previous ones. At this point, we can either stop accumulating frames, and extract the activity area, or we can gather more frames, to extract a region corresponding to additional motions. In Fig. 1(a) we show the Normal Probability Plot (NPP) of a background pixel's flow estimates (for the Tennis video in Sec. 5) accumulated over 50 frames, which shows that it indeed follows an approximately Gaussian distribution. Similarly, Fig. 1(b) shows the NPP for the flow estimates of a pixel that belongs to the moving object: as expected, in this case the random variable (the flow estimates) deviates from the Gaussian model, due to the actual motion taking place.

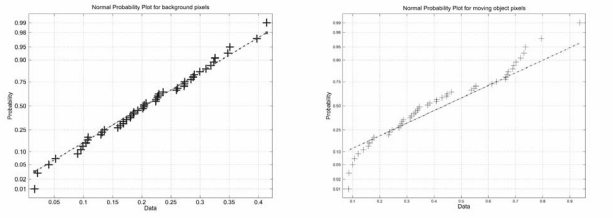

Fig. 1. Normal Probability Plots for flow estimates corresponding to: (a) a background pixel, (b) a moving pixel.

We then estimate the kurtosis of each pixel's velocity estimates over the frames examined. Pixels with non-zero kurtosis have been displaced during the frames being examined, and pixels with zero kurtosis correspond to pixels with no motion. In practice, we have a finite number of velocity estimates, so they do not follow any distribution perfectly, and consequently do not achieve exactly zero kurtosis in the non-activity areas. Thus, we consider the pixels with kurtosis less than 10% of the mean kurtosis to be immobile. Our experiments show that this is indeed a reliable way to extract the activity areas, which is also generally applicable to any collection of motion estimates. In Fig. 2(b)-(d) we show the results of accumulating the flow over 10, 20 and 30 frames, where the signatures of the actions taking place in each subsequence become evident. It should be noted that the number of frames used for the activity areas does not affect the final segmentation results, since these regions will be a superset of the pixels occupied by the moving objects.

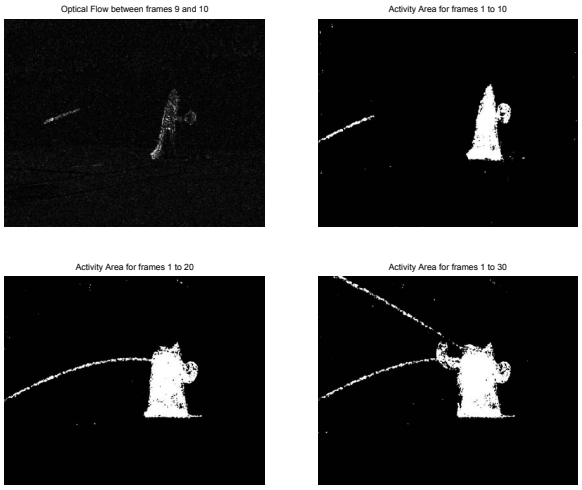

Fig. 2. Tennis game. (a) Optical flow. Activity areas for (b) frames 1 to 10, (c) frames 1 to 20, (d) frames 1 to 30.

3. COLOR AND MOTION FUSION

The more accurate localization of the moving objects in a video can make even richer semantic interpretations possible, e.g. by extracting the actual object trajectories, along with the objects themselves. For this reason, we further process the activity areas previously extracted (Sec. 2), along with color information, by performing color segmentation in the background and the activity areas. The colors are clustered using the mean-shift algorithm, as it does not require the determination of the number of clusters in each frame, and has been shown to give reliable segmentation results in practice. Mean shift models the data's distribution by a kernel $K(\mathbf{x})$, as follows:

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right), \quad (3)$$

where $\mathbf{x}$ is the $d-$dimensional data and n the number of data samples, and h the search radius. The algorithm then searches for this distribution's modes iteratively. We use the Epanechnikov kernel, shown below, as it is symmetric and differentiable, enabling us to find the distribution's gradient, and then its modes.

$$K_E(\mathbf{x}) = \begin{cases} \frac{1}{2} c_d^{-1}(d+2)(1 - \mathbf{x}^T\mathbf{x}), \; if \mathbf{x}^T\mathbf{x} < 1 \\ 0, \; \mathbf{otherwise} \end{cases} \quad (4)$$

In [8], it is shown that the modes of this distribution are found by successively translating the data window by the "sample mean shift", given by

$$M_h(\mathbf{x}) = \frac{1}{n_x} \sum_{\mathbf{x}_i \in S_h} \mathbf{x}_i - \mathbf{x}, \quad (5)$$

where n_x is the number of samples in the search area S_h. The pixels whose color is closest to the density maxima derived by the mean shift are assigned to those cluster centers. The color information alone provides a cue about the semantics, as it can indicate the kind of the tennis court (e.g. in Fig. 4 we have clay and not grass).

In order to isolate the moving entities in the video, we apply mean shift color segmentation to the background areas, as well as the activity ares in each frame. This leads to several layers of color, in both regions. We then compare the (color) histograms of each color layer using the Earth Mover's Distance (EMD), which computes the distance between two distributions. The EMD [10] is defined as the minimum amount of work needed to change one distribution (signature) into the other, where "work" refers to the user-defined ground distance between two features; in our case, the Euclidean distance is employed to this end. The signatures involved in the computation of the EMD are defined as: $S = \{s_j\} = \{(\mathbf{m}_j, w_j)\}$, where $\mathbf{m}_j$ represents a d-dimensional point (here, the three mean color values of each histogram bin) and w_j is the corresponding histogram value. Then, each color layer of the activity area is matched to the color layer of the background area for which it has the smallest EMD. Since the color of the moving object's pixels differs the most from the background layers' colors, this approach successfully isolates the moving object in the activity area of each frame, with very good results, as shown in Sec. 5.

4. SEMANTICS EXTRACTION

The usefulness of the extracted activity areas is not restricted to the accurate localization of the moving objects. The activity areas are often characterized by a shape representative of specific actions, so they can aid in the extraction of semantic information concerning the sequence. In Fig. 2(a) we show representative optical flow estimates between frames 9 and 10 of a tennis game sequence. As expected, they have higher values at the borders of the moving objects, but also contain noise terms in pixels that did not move.

In Fig. 3 we show activity areas for a tennis serve, after the accumulation of different numbers of frames. As before, we can see the different characteristics of the actions taking place in each subsequence. In Fig. 3(a) there is a large curve, caused by the racket hitting the ball, which is very characteristic of a tennis serve, whereas Fig. 3(b) has a curve on the bottom left, corresponding to the tennis player pulling the racket back after she hit the ball. In both cases, we also see the symmetric signature of the player's leg motion on the right. By comparing these results for a tennis serve with the activity areas of the tennis game in Fig. 2, we can easily distinguish them, and immediately understand which action is taking place in each case. The differentiation of a tennis serve from simply hitting the ball is particularly useful in real applications, since such semantics are very commonly utilized in the training of tennis players, which is often based on the analysis of videos.

For an actual video application, this distinction can be automatically evaluated using shape descriptors, such as those defined by the MPEG-7 Standard [11]. We describe the shapes of the activity areas with a 2D contour-based shape descriptor, because the information that is most revealing about the activities being performed is contained in the contours of the activity areas. We extracted these descriptors for the activity areas corresponding to the tennis game and the tennis serve, and, as expected, they differentiate effectively between the two actions. This representation allows the integration of the motion processing stage (Sec. 2) in a fully automated video semantics extraction system, which could search for similar activities in a video by comparing the MPEG-7 shape descriptors of the contours of their respective activity areas.

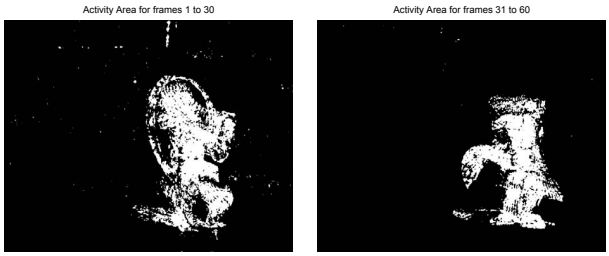

Fig. 3. Tennis serve activity areas. Frames: (a) $1-30$, (b) $31-60$.

5. EXPERIMENTS

The above developed methodologies were applied to real heterogeneous videos. In the sequel, results are shown for sports applications, for two different sports, namely Tennis and Formula 1 racing.

Tennis: A sequence with a tennis player serving the ball is examined (Fig. 4(a)). The motion estimation (in C++) stage took 0.4 seconds per frame, and the mean shift color segmentation (in Matlab) 2.4 seconds per frame, on a dual core Pentium 4, 3.4 GHz processor. The optical flow and several activity areas for it are shown in Fig. 2. The activity areas can be used to interpret what is happening in the scene, namely a tennis serve, as previously discussed. By applying the mean shift color segmentation to the background and activity areas, we also obtain layers of color that are consistent with what we expect (Fig. 4(b)-(c)). Namely, the color layers of the moving object in the activity area deviate significantly from the corresponding layers in the background. This difference can then be reliably used to segment the moving object, and indeed gives good results, as seen in Fig. 4(d). The color segmentation results combined with the motion information provide us with a complete segmentation of the tennis player.

Formula 1 video: Experiments are also performed with a Formula 1 video; here, indicative results for a shot featuring a series of cars turning the bend in front of the camera (Fig. 6(a)) are shown. The motion estimation (in C++) took 0.39 seconds per frame, and the mean shift color segmentation (in Matlab) 3 seconds per frame. The optical flow estimates are higher near the borders of the cars, but there is also illumination noise, which introduces erroneous estimates in the background areas (Fig. 6(b)). Despite these errors, and the poor quality of the video, the method of Sec. 2 succeeds in extracting the activity area in the area where the cars are moving during the video, shown in Fig. 6(c). Similarly to the tennis serve, this area can also be used to extract semantic information, e.g. that the video being processed is being filmed in the race track in front of the turning point, and that the cars have started and are turning the bend. The color segmentation of the background and activity areas in a representative frame are shown in Fig. 6(d)-(e). We see that the moving objects' colors in the activity area are separated from those of the background, so when the color layers are compared, we expect to separate the cars from the rest of the scene. Indeed, Fig. 6(f) shows that the comparison of the color histograms led to the successful isolation of cars turning in frame 20.

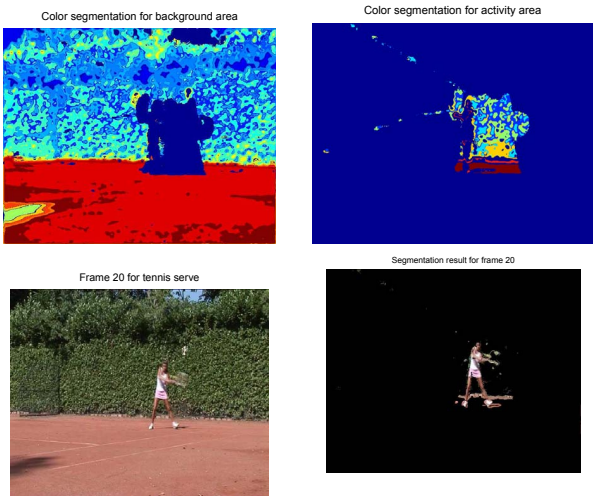

Fig. 4. Frame 20 of Tennis. Color segmentation for (a) background areas, (b) action areas. (c) Frame 20. (d) Final segmentation.

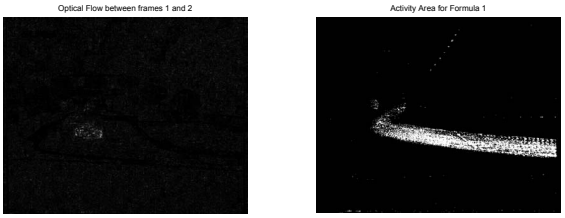

Fig. 5. Formula 1 video. (a) Optical flow. (b) Activity area.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a novel approach for the derivation of higher level information from video sequences. We extract areas of activity in a video, by accumulating the flow fields over several frames and processing their statistics. The integration of the resulting activity areas with color provides valuable information about the semantics of the video, such as the actions taking place and the moving entities detected in it. The extracted objects and events could even be incorporated in the future in a system that searches for similar objects in a database, in other videos or video segments.

Acknowledgments: This work was supported by the European Commission under contracts FP6-001765 aceMedia, FP6-027685 MESH and FP6-027026 K-Space and by the GSRT funded project DELTIO: Analysis of Multimedia Content using Evolutionary Ontologies and Application to Television News Bulletins.

7. REFERENCES

[1] C. Dorai et al., "Media semantics: who needs it and why?," in *Proc. ACM Multimedia*, Dec. 2002, p. 580583.

[2] Hwang J. and Luo Y., "Automatic object-based video analysis and interpretation: a step toward systematic video understanding," in *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, May 2002, vol. 4, pp. IV–4084 – IV–4087.

[3] N. Dimitrova and F. Golshani, "Motion recovery for video content classification," in *ACM Trans. Information Systems*, Oct. 1995, vol. 13, p. 408439.

[4] Naphade M. R., Kozintsev I.V., and T.S. Huang, "A factor graph framework for semantic video indexing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 1, pp. 4052, Jan. 2002.

[5] Dasiopoulou S., Mezaris V., Papastathis V.K., Kompatsiaris I., and Strintzis M.G., "Knowledge-assisted semantic video object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1210–1224, 2005.

[6] Sadlier D.A. and N.E. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 10, pp. 1225 1233, 2005.

[7] Lukas B. and Kanade T., "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[8] Comaniciu V. and Meer P., "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 603 – 619, May 2002.

[9] G.B. Giannakis and M. K. Tsatsanis, "Time-domain tests for gaussianity and time-reversibility," *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3460 – 3472, Dec. 1994.

[10] Rubner Y., Tomasi C., and Guibas L. J., "The earth movers distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99 – 121, 2000.

[11] Bober M., "Mpeg-7 visual shape descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 716 – 719, June 2001.

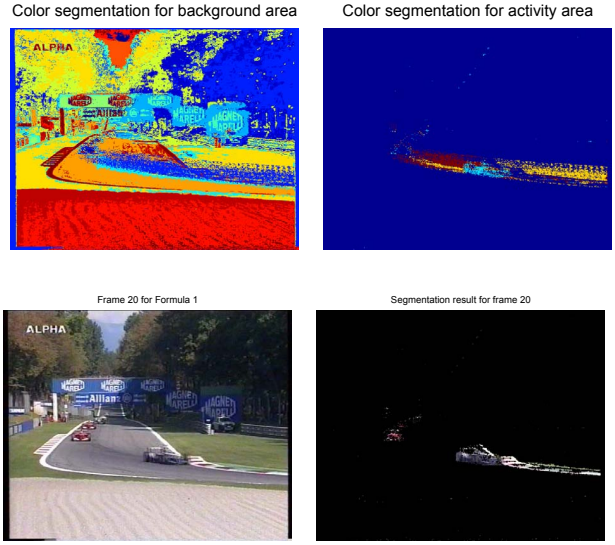

Fig. 6. Frame 20 of Formula 1. Color segmentation for (a) background areas, (b) action areas. (c) Frame 20. (d) Final segmentation.