Video Content Representation as Salient Regions of Activity

Nicolas Moënné-Loccoz, Eric Bruno, and Stéphane Marchand-Maillet

University of Geneva 24, rue General Dufour - 1211 Geneva 4, Switzerland

{Nicolas.Moënné-Loccoz, Eric.Bruno,
Stephane.Marchand-Maillet}@cui.unige.ch,
<http://viper.unige.ch/>

Abstract. In this paper, we present a generic and robust representation of video shots content expressed in terms of salient regions of activity. The proposed approach is based on salient points of the image space, thus minimizing the computational effort. Salient points are extracted from each frame. Their trajectories are computed between successive frames and the global motion model is estimated. Moving salient points are selected from which salient regions are estimated using an adaptive Mean-Shift process, based on the statistical properties of the point neighborhoods. The salient regions are then matched along the stream, using the salient points trajectories. The information carried by the proposed salient regions of activity is evaluated and we show that such a representation of the content forms suitable input for video content interpretation algorithms.

1 Introduction

Video content representation is the task of extracting information from a video stream in order to facilitate its interpretation. We base our study on video shots as they are temporal segments of the video with consistent visual content. The information carried by video shots is twofold. First, a video shot holds a visual context, i.e. the visual environment and the editing effects. A video shot also contains events, i.e. long-term temporal objects [16]. We focus our approach on the extraction of such events and aim at a well defined representation of the visual content.

To extract events occurring in video shots, most approaches either rely on domain-specific algorithms or on global features. Works using local features in an unconstrained capturing environment are mainly based on spatio-temporal segmentation of the entire scene, which is a difficult problem that has no generic solution. As the activity of a scene is limited to a given number of regions, we propose to extract salient regions of activity. By definition, these are salient regions of the scene having homogenous perception properties and a salient temporal behavior. This paper presents a robust and fast process to extract salient regions of activity derived from spatio-temporal salient points. A representation

of the video content is obtained as the set of the salient regions defined in the 2D+T space of the video stream. Salient regions of activity highly correlate with the events occurring in the considered scene. As such, they allow the detection of events that can be used by video content interpretation algorithms.

Section 2 describes the spatio-temporal salient points extraction process. Multi-scale salient points are first extracted from each frame of the video stream. Then, salient points are matched between successive frames in order to estimate their trajectory. The global affine motion model is estimated from the trajectories of the points. Spatio-temporal salient points are selected as the points that do not follow the global motion. Section 3 details how salient regions are estimated from moving salient points using an Adaptive Mean-Shift process that is based on the color distribution of the point neighborhoods. The extracted salient regions of activity are matched along the stream by considering the trajectories of the points belonging to it. Section 4 provides an evaluation of the information carried by such a representation, demonstrating its robustness and validating its usefulness in the context of video content interpretation. Finally section 5 discusses the main contributions of the present work.

2 Spatio-temporal salient points

2.1 Spatial Salient points

Salient points are points in the image space where the intensity changes in at least two directions ; because they carry high information about the structural content of a scene, they are used for image retrieval [4, 14], imaging parameters estimation [1, 17] and objects recognition [8]. Several algorithms have been proposed to extract salient points such as the *Plessey feature points* [5], *SUSAN* [12], *Curvature Scale Space* [10] and the Wavelet based [6] approach.

K. Mikolajczyk and C. Schmid in [9] present a multi-scale version of the Plessey feature points. Their points are among the most robust and have the desirable property to be scale invariant. The multi-scale interest points are extracted from the scale space of an intensity image $I(v), v \in V = \{x, y\}$. The scale-space is obtained by convolving $I(v)$ with a Gaussian derivative kernel for a set of scales $s \in S$:

$$L_{v_i}(v, s) = I(v) \star G_{v_i}(s), \forall v \in V, \forall s \in S \quad (1)$$

where $G_{v_i}, i \in \{1, 2\}$ is the Gaussian derivative along the dimension v_i of the image space V .

The scale normalized Harris function $H(v, s)$ is computed at each scale s and each location v :

$$H(v, s) = Det(\Sigma(v, s)) - \alpha Trace^2(\Sigma(v, s)) \quad (2)$$

$$\Sigma(v, s) = s^2 G(v, \sigma) \star \begin{pmatrix} L_{v_1}^2(v, s) & L_{v_1} L_{v_2}(v, s) \\ L_{v_1} L_{v_2}(v, s) & L_{v_2}^2(v, s) \end{pmatrix} \quad (3)$$

$H(v, s)$ gives a measure of the cornerness of the points v at the scale s . It is based on the strength of the eigenvalues of the auto-covariance matrix Σ of L_v .

According to the scale-space theory [7], the characteristic scale of a point is a local maxima of the Laplacian defined by :

$$Lp(v, s) = s^2 |L_{v_1 v_1}(v, s) + L_{v_2 v_2}(v, s)| \quad (4)$$

Thus, multi-scale salient points are defined to be points that are local maxima of H in the image space and local maxima of L in the scale space.

Such interest points defined by K. Mikolajczyk and C. Schmid are extracted for each frame of the video stream. In the sequel, we note W_t the set of salient points w extracted independently at the frame F_t and s_w the characteristic scale of point w .

2.2 Temporal salient points

Spatial salient points are located on the objects of interest but also on textured surfaces or non-informative background structures. For these reasons, we consider only the moving salient points, i.e. points which motion is different from that of the background. To extract these temporal salient points, trajectories of the spatial salient points are computed and the background motion model is estimated. Temporal salient points are selected as the spatial salient points having a salient temporal behavior for a given period of time.

Salient points trajectories The trajectories of salient points are computed by matching them between two consecutive frames. More formally, for any $w_t \in W_t$, the corresponding point $w_{t-1} \in W_{t-1}$ is selected. Points in W_t and W_{t-1} are first described by a the 9-dimension vector of the Hilbert invariants up to the third order as presented by C. Schmid in [11]. Then, the Mahalanobis distance $d(w_t, w_{t-1})$ is computed for each pair of points. To select the good matches M_t among the set of possible matches, a greedy algorithm is applied that tends to minimize the sum of the distances $\sum_{(w_i, w_j) \in M_t} d(w_i, w_j)$.

We therefore obtain the set of matches M_t that associates the salient points of W_t in the current frame with their predecessors in the previous one. The set of match M_t corresponds to the set of trajectories of the points between the two successive frames F_{t-1} and F_t .

Global motion model estimation Given the set of trajectories M_t , we estimate the most representative affine motion model of the trajectories (see [13] for an overview of motion estimation). This model thus corresponds to a global description of the background motion. We choose the affine motion model because of its ability to capture the main camera motions with a limited number of parameters:

$$d(v) = \begin{pmatrix} a_1 & a_2 \\ a_4 & a_5 \end{pmatrix} v + \begin{pmatrix} a_3 \\ a_6 \end{pmatrix}$$

To estimate the motion model from the set of trajectories M_t , we first apply a RanSaC algorithm [3] which tends to select the most representative motion model. As the set of trajectories contains noise, the model is then smoothed by applying a *Tukey M*-estimator in way close to the one presented in [15].

Points selection Moving salient points are those that do not follow the background motion model. In order to remove potential noise, only salient points detected as moving for a given time interval are selected (in our experiments we set this interval to 10 frames). Such points are mainly situated on moving objects and provide high information about the dynamic content of the scene. However, as they are located on corner like regions, they do not well focus on the visual events.

3 Salient Regions of activity

Salient regions of activity are homogenous (in terms of perception properties, e.g. color distribution) moving regions defined in the 2D+T space of the video shot. They tend to correspond to the events (as defined in [16]) occurring in the scene. We detect and track such regions from the moving salient points. Hence, they inherit the saliency and the robustness of such features. We model salient regions of activity as a succession in time of spatial ellipses. The parameters of the ellipses are estimated by a *Mean Shift* algorithm augmented with a kernel adaptation step.

3.1 Adaptive Mean Shift algorithm

The Mean Shift algorithm has been used to track regions of interest [2]. The main idea is to compute an offset δ_{v_r} between a current estimation of the region location v_r and an estimation v'_r having a higher likelihood. The offset is computed by :

$$\delta_r = \frac{\sum_v K(v - v_r)p(v)(v - v_r)}{\sum_v K(v - v_r)p(v)}$$

where K is a kernel centered in v_r and $p(v)$ is a weighting function measuring the probability of the pixel v to belong to the region. From an initial location, the algorithm computes a new location from the offset δv_r and is iterated until the location converges to a local maxima.

In our setting, we use an ellipsoidal Epanechnikov Kernel because it has a convergence rate much more higher than the Gaussian Kernel :

$$K(v - r) = \begin{cases} \frac{3}{4} \left(1 - \left(\frac{(v-v_r)^T \Sigma_k (v-v_r)}{\sigma_s} \right)^2 \right) & \text{if } \left| \frac{(v-v_r)^T \Sigma_k (v-v_r)}{\sigma_s} \right| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

where Σ_k is the affine transformation corresponding to the shape of the kernel and σ_s is the size of the kernel (its radius in the affine transformed domain).

As the shape and the size of the salient regions of activity are not known a priori, a shape adaptation step is necessary after the convergence of the Mean Shift. The shape and the size of the kernel are estimated from the covariance matrix of $p(v)$.

The algorithm then alternates a Mean Shift algorithm with an adaptation step until both the location of the region v_r and the kernel converge. In order to speed up the process, some divergence criteria may be defined such as a maximum size of the Kernel and a minimum sum of the weights $p(v)$.

3.2 Region detection

A new salient region of activity is detected for each moving salient point not within a currently tracked region. The initial position of the region is set to the position of the corresponding salient point : $v_r = w$. The shape of the kernel is initialized to a circle, i.e. $\Sigma_k = \text{diag}(1, 1)$ and its size to its characteristic scale, i.e. $\sigma_s = 3 * w_s$.

As weighting function, we use $P(v|\theta_w) = N(\mu_{\theta_w}, \Sigma_{\theta_w})$, the probability of a pixel to be in the Gaussian *RGB* normalized ($R/(R + G + B), G/(R + G + B)$) color space defined by θ_w . The parameters θ_w are estimated in the spatial neighborhood of the salient point (i.e., the neighborhood inside the circle of radius $3 * w_s$ centered at the point location).

An adaptive mean shift process is then performed that estimates the ellipse maximizing the likelihood of the region.

3.3 Region tracking

Detected salient regions of activity are tracked along the shot. First, the position of a region is updated by adding to its previous position the mean motion vector of the salient points within it : $v_r^{t+1} = v_r^t + E[(w_{t+1} - w_t)]$. Then, an adaptive mean shift process is performed which updates the ellipse estimate.

Figure 1 presents examples of detected and tracked salient regions in different settings.

4 Evaluation

In order to evaluate the information provided by the extracted salient regions of activity we consider the events they characterize and evaluate the *precision* of this information. This study is performed over a corpus containing about 50 shots: 22 *sport* shots, 13 *news* shots, 14 *sitcom* shots (from the *MPEG7 evaluation corpus*) and 5 videos of meeting (from the *M4 meeting corpus*).

We have labeled about 1000 extracted salient regions according to the events they characterize (see table 1). We observe a precision above **63%**. In other words, almost 2 out of 3 regions are semantically meaningful. Hence, salient regions of activity are well-suited to provide video interpretation algorithms with a meaningful decomposition of the spatio-temporal content in terms of events.

Number of regions	Label
312	Moving Human Body
122	Moving Head
56	Moving Arm
62	Moving Leg
24	Agitated Group of Person
45	Text Overlay
9	Moving Ball
364	Noise
994	Total

Table 1. Salient regions labels.

Table 2 shows a clear correlation between the labels of the detected salient regions of activity and the class of the video shot. For example, salient regions extracted in *sport action* sequences, correspond mainly to *Moving Human body* and *Moving Leg* (running player). For *Speaking Anchor Person* sequences, regions mostly correspond to *Moving Head* and *Text Overlay* (anchor person and textual information). Thus, these results show that salient regions of activity are consistent with the content of the shot and form cues that could be used to perform event retrieval or recognition.

	Sport Action	Speaking Anchor Person	Meeting Action
Moving Human Body	39.09	9.60	25.82
Moving Head	4.2	27.40	30.05
Moving Arm	3.87	12.80	8.45
Moving Leg	10.01	0.00	0.00
Agitated Group of Person	3.87	0.00	0.00
Text Overlay	4.36	16.10	0.00
Moving Ball	1.45	0.00	0.00
Noise	33.11	33.80	35.68
Total	100	100	100

Table 2. Labels repartition according to the shot class.

The *recall* of this representation cannot be evaluated in a systematic way because it calls for the exhaustive annotation of all events (defined as long-term temporal objects) of the corpus performed by naive subjects. However, a systematic visual inspection of the spatio-temporal location of the detected regions let us envisage that they highlight most of the events and that they could be used by higher-level algorithms as a focus of attention preprocessor. Figure 1 shows some visual illustrative examples.

5 Conclusion

In this paper, we have presented a novel representation of video content that captures the events occurring in the stream. The proposed procedure for salient regions of activity extraction is fast (less than 1 second per frame in our matlab implementation) robust (based on invariant features) and generic (it makes no assumption on the type of content at hand).

Our results clearly demonstrate that this representation is coherent (i.e. robust and meaningful) and is an effective decomposition of the content that can be further exploited by interpretation processes.

Future work will therefore focus on the use of such a representation to develop new event retrieval algorithms in videos databases.

Acknowledgments

This work is funded by EU-IST project M4 (www.m4project.org) and the Swiss NCCR IM2 (Interactive Multimodal Information Management).

References

1. A. Baumberg. Reliable feature matching across widely separated views. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 774–781, 2000.
2. D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. pages 142–151, 2000.
3. M.A. Fisher and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In *Communications of the ACM*, pages 381–395, 1981.
4. V. Gouet and N. Boujemaa. On the robustness of color points of interest for image retrieval. In *IEEE International Conference on Image Processing ICIP'2002*, 2002.
5. C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pages 189–192, 1988.
6. J.-S. Lee, Y.-N. Sun, and C.-H. Chen. Multiscale corner detection by using wavelet transform. In *IEEE Transactions on Image Processing*, 1995.
7. T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116, 1998.
8. David G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu*, pages 1150–1157, 1999.
9. Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *8th International Conference on Computer Vision*, pages 525–531, 2001.
10. F. Mokhtarian and R. Suomela. Robust image corner detection through curvature scale space. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, 20(12):1376–1381, 1998.
11. C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, 19(5), 1997.

12. S. M. Smith and J. M. Brady. SUSAN – A new approach to low level image processing. Technical Report TR95SMS1c, Chertsey, Surrey, UK, 1995.
13. C. Stiller and J. Konrad. Estimating motion in image sequences: A tutorial on modeling and computation of 2D motion. *IEEE Signal Process*, 16:70–91, 1999.
14. Q. Tian, N. Sebe, M. S. Lew, E. Loupiau, and T. S. Huang. Image retrieval using wavelet-based salient points. In *Journal of Electronic Imaging, Special Issue on Storage and Retrieval of Digital Media*, pages 835–849, 2001.
15. Philip H. S. Torr and Andrew Zisserman. Feature based methods for structure and motion estimation. In *Workshop on Vision Algorithms*, pages 278–294, 1999.
16. L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.
17. Zhengyou Zhang, Rachid Deriche, Olivier D. Faugeras, and Quang-Tuan Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1-2):87–119, 1995.

Meeting room (static camera)

News (static camera)

News (moving camera)

Basketball game (moving camera)

Soccer game (moving camera)

Fig. 1. Samples of Salient Regions of Activity.