

SINGLETS: Multi-Resolution Motion Singularities for Soccer Video Abstraction

Katy Blanc, Diane Lingrand, Frédéric Precioso

▶ **To cite this version:**

Katy Blanc, Diane Lingrand, Frédéric Precioso. SINGLETS: Multi-Resolution Motion Singularities for Soccer Video Abstraction. Workshop CVsports (in conjunction with CVPR), Jul 2017, Honolulu (Hawaii), United States. hal-01540342

HAL Id: hal-01540342
https://hal.science/hal-01540342v1

Submitted on 1 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

SINGLETS: MULTI-RESOLUTION MOTION SINGULARITIES FOR SOCCER VIDEO ABSTRACTION

Blanc Katy Lingrand Diane Precioso Frederic
Université Côte d'Azur, CNRS, I3S, France
{kblanc, lingrand, precioso}@i3s.unice.fr

Abstract

The burst of video production appeals for new browsing frameworks. Chiefly in sports, TV companies have years of recorded match archives to exploit and sports fans are looking for replay, summary or collection of events.

In this work, we design a new multi-resolution motion feature for video abstraction. This descriptor is based on optical flow singularities tracked along the video. We use these singlets in order to detect zooms, slow-motions and salient moments in soccer games and finally to produce an automatic summarization of a game.

We produce a database for soccer video summarization composed of 4 soccer matches from HDTV games for the FIFA world cup 2014 annotated with goals, fouls, corners and salient moments to make a summary. We correctly detect 88.2% of saliant moments using this database. To highlight the generalization of our approach, we test our system on the final game of the handball world championship 2015 without any retraining, refining or adaptation.

1. Introduction

In the world of digitization we live in, big quantities of numerical data are stored and available. This is particularly common for image and video databases since a simple clip in high definition of 30s contains 750 frames (with 25fps) and therefore more than a billion pixels. The amount of stored videos keeps increasing every hour. For instance, last IPhone and Samsung smart phones take short videos instead of static pictures, that they called respectively "live photos" and "motion photos". "Live photo" is perfectly chosen name since even a very brief motion recording gives life to a simple picture. Besides every Facebook user can now have a video profile on his main own page instead of a classic profile picture. The number of Youtube videos increased to such an extent that Google Research just released a dataset called YouTube-8M that contains half a million hours of videos [1]. To face this always increasing amount of videos, algorithms must analyze and recognize the content of these elements to be able to extract knowledge, statistics and everything that a user could look for. With over a decade of extensive research, there has been a tremendous development in the domain of video content mining [13, 24].

In this article, we present a new motion descriptor based on extracting singularities in the motion domain in Section 3, that is detecting specific motion patterns in the optical flow. In Section 3.2, we build a unified description of optical flow that allows us to describe different aspects of video semantics. We call this descriptor singlet which corresponds to motion singularities at different resolutions tracked along a video. As a good context of video abstraction, we focus on the application of our new motion features for sport analysis, more specifically soccer matches. In the experiment section, our flow description is detailed for zoom detection, extraction of salient moments and slow-motion detection. We then propose an automatic method of production of video skimming. Finally, we evaluate the singlets' efficiency in various contexts: soccer games from the World cup 2014 and the final game of handball world championship 2015.

2. Related Works

We classify the contributions regarding video analysis along two axes: retrieval and abstraction [10].

On the one hand, video retrieval is a very active research field aiming at providing tools to retrieve videos from content. The current state of the art methods for video content representation start by detecting keypoints. Most of the time, these keypoints are then tracked along the video, and finally described by combining a description of both static and dynamic visual information (shape, appearance...). Among these contributions, let us cite the seminal work on Space-Time Interest Points (STIP) [12] which proposed to combine Histograms-on-Oriented-Gradients (HOG) with Histograms-on-Optical-Flow (HOF). In [17], the tracklets are also defined from HOG and HOF and classified using dynamic time warping. Improved dense trajectories (iDT) adds SURF and MBH (derivative of HOF) to the description, keeping a combination of keypoint displacements [25].

However all these standard motion descriptors usually do not hold complex motion information. Recently, deep learning convolution networks for video description, such as the C3D network [22], have been considered, but for now they only achieves complementary results compared to iDT.

Abstraction consists in segmenting the video stream into consistent content units and generating a brief overview of the video from these units. Truong *et al.* [23] lists and compares techniques providing video abstraction by distinguishing video summarization, made of static video keyframes, from video skimming, made of dynamic video shots. Four types of video are much analyzed: meetings, movies, broadcast news and sports [28]. The growing mass of available video data is well illustrated in the sports domain: TV channels have broadcast matches in soccer competitions for years. Nowadays, the amount of sport broadcast has drastically increased with the legalization of sports betting. The algorithms to enhance these broadcasts are eagerly awaited and related research works have emerged.

In video abstraction and in particular in soccer video abstraction, people do not use the state of the art representation, iDT, which is made for retrieval and is not suitable for motion abstraction. From a single match, there is already a big amount of information to extract in order to summarize the match: number of passes or goals, players' statistics... For example, by focusing on play-break session can filter video information [6]. To detect the play-break moments, Xu *et al.* [29] used heuristic rules based on view classification (long, medium and close-up views) while Xie *et al.* [27] used a stochastic structured model by HMM. Also in a stochastic approach, Leonardi *et al.* [14] proposed controlled Markov chains to detect goals.

In general, the first approach in soccer video analysis is to segment shots and extract features. These low-level features are extracted from the frames and are often associated with hand-crafted features, like line marks, ball tracking or the color of the players jersey. Gong *et al.* [8] use these features to classify events like shots or corner kicks.

A soccer match is quite long and contains few interesting actions, which is why detecting salient moments is the most active research side in the soccer video domain. Other hand-crafted features like the overall excitement computed from shot length, motion and audio activity are used to characterize game extracts in [9]. These features are found by building mid-level representation of the shots that can be learned to recognize events: Duan *et al.* [4] used shot lengths and texture maps to train a SVM while Wu *et al.* [26] used global motion estimation to train a neural network. Sadlier *et al.* [19] focused on audio features, and low-level features to train a SVM and they claims their techniques to be generic while only focusing on "field sports" and thus benefiting from knowing the ground color (grass), field line marks and others shared characteristics. Yow *et al.* [31] built a panoramic representation of a shot and then select important shots using heuristics. Assfalg *et al.* [2] and Tabii *et al.* [21] used finite state machines, respectively one based on ball motion, players positions and jerseys and the other one based on play field segmentation, shot detection and classification, however this method requires good handmade rules. Ye *et al.* [30] chose an incremental learning SVM on a mid-level description containing the features of successive moments like view labels, line mark positions and shot descriptors to select highlights. After using an artificial neural network in order to detect the appearance of logos and score and using Hough and k-means to detect goal mouths, Zawbaa *et al.* [32] classified goals, attack and other events with an SVM from the detection of goal mouth position. In recent work, Raventos *et al.* [18] ranked shots and their associated keyframe using face and skin detection, whistle detector and user specifications.

All these methods use at best an energy function to measure the motion activity whereas the motion in a soccer game is probably the most important information. We now present our adaptive and generic representation to describe the motion and show how this representation allows us to build a video abstraction.

3. Video movement analysis

Our method of motion content analysis is inspired by the work of Kihl *et al.* [11] which extracts singularities from motion in the domain of fluid mechanics. A singularity is a vanishing point akin to root for polynomials. It can be seen like keypoints in optical flow. Optical flows are projected to polynomial bivariable function space to detect these vanishing points from their polynomial approximations. Starting from the definition of these singularities, we design a new local motion video content descriptor.

3.1. Polynomial projection of optical flow and singularities

The two horizontal and vertical components of the optical flow U and V at each pixels (x_1, x_2) are computed using the method of Gunnar Farneback [7]. U and V are projected onto Legendre basis to get the best approximation in the polynomial optical flow space. Then they are expressed in a canonical basis.

$$P_U(x,y) = \sum_{k=0}^{K}\sum_{l=0}^{L} u_{k,l}.x^k.y^l; P_V(x,y) = \sum_{k=0}^{K}\sum_{l=0}^{L} v_{k,l}.x^k.y^l \quad (1)$$

with $K + L \leq degree$. Similarly to the work of Kihl *et al.* [11], we restrict approximations to degree 1.

$$\begin{pmatrix} U \\ V \end{pmatrix} \simeq \mathbf{A}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \mathbf{b} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + b_1 \\ a_{21}x_1 + a_{22}x_2 + b_2 \end{pmatrix} \quad (2)$$

Indeed, according to $\mathbf{A}$ and $\mathbf{b}$, the singularities of the optical flow appears at the position $(x_1\, x_2)^T = -\mathbf{A}^{-1}\mathbf{b}$. The type of a singularity depends on $tr(A)$ and $\Delta(\mathbf{A})$.

$$\Delta(\mathbf{A}) = (tr(\mathbf{A}))^2 - 4\,\det(\mathbf{A}) \tag{3}$$

We present the different configurations in figure 1.

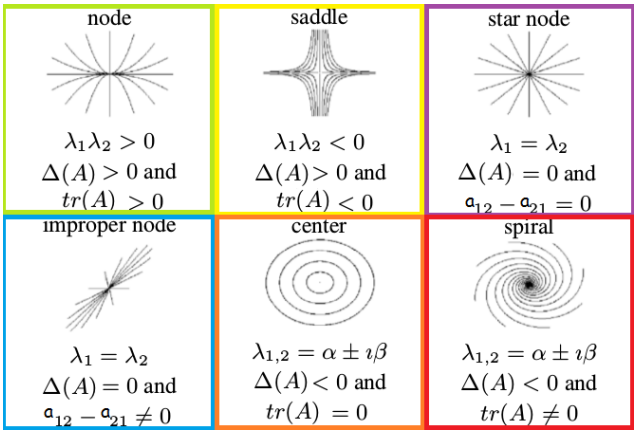


Figure 1. Classification of singularities based on the value of $\mathbf{A}$ (illustration from [11]).

Singularities are detected using a sliding window at different scales (from $0.1h$ to $0.5h$ by steps of $0.1h$ where h is the smallest dimension of the video, usually the height). This makes the singularity to be multi-resolution compliant. Multiple singularities are detected at different scales at the same position. These multiple singularities have not been merged, as for SIFT features, since they could be of different types or intensity and carry different information.

By definition, a singularity corresponds to vanishing points. Thus no singularity will be detected in a pure translation. A singularity is not detected in situations where $\mathbf{A}$ is not invertible. Moreover the highest the determinant is, the more significant singularities are preserved: a threshold on $\det(\mathbf{A})$ permits to filter the noisy singularities and to keep informative ones. No thresholding on $\det(\mathbf{A})$ implies to keep every singularities, even the ones with very small motion which are most probably produced by pixel noise.

Thereafter, we globally or locally extract these singularities, according to what we focus on.

3.2. Spatio-temporal representation

As introduced above, singularities can be tracked along a video to describe a time lapse.

For T frames, $T-1$ optical flows are computed. On each of these optical flows, we extract the singularities. In order to build chains of singularities, called *singlets*, they are matched in reverse time. Based on the hypothesis that a singularity has a small displacement between two frames and to reduce computational cost, we match only singularities in a near neighborhood. Thus, for each singularity $sing_s$ in the optical flow f_t, the singularity candidates in the optical flow f_{t-1} are restricted to a near neighborhood $V(sing_s)$. Two singularities are considered in the same neighborhood if their sliding windows have a suitable overlap ratio (figure 3.2) as defined directly below:

$$V(sing_s) = \left\{ sing; \frac{area(W(sing) \cap W(sing_s))}{area(W(sing) \cup W(sing_s))} > \alpha \right\} \tag{4}$$

where $W(sing_s)$ is the sliding window of $sing_s$.

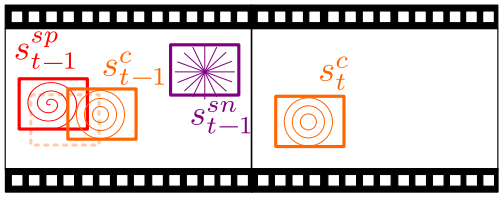


Figure 2. Two successive optical flows: Searching a match for the singularity s^c_t within the singularities in the previous optical flow. Regarding the overlap ratio, s^{sp}_{t-1} and s^{c}_{t-1} are candidates while s^{sn}_{t-1} is not. The match is the closest singularity regarding localization and type: s^{sp}_{t-1}. Singularity types sp, sn and c respectively correspond to spiral, star node and center.

To match singularities during time, the best candidate within a close neighborhood is selected as the one which minimizes the singularity distance described below:

$$d(\begin{pmatrix}\mathbf{A}\\ x\\ y\end{pmatrix}, \begin{pmatrix}\mathbf{A}'\\ x'\\ y'\end{pmatrix}) = ||A-A'||_F + \lambda||\begin{pmatrix}x\\ y\end{pmatrix} - \begin{pmatrix}x'\\ y'\end{pmatrix}||_2 \tag{5}$$

where λ is a weight balancing the distance between the positions and the similarity between singularity coefficients values. In each affine optical flow f_t, a singularity $sing_s$ is described by 6 coefficients, 4 in $\mathbf{A}$ and 2 in $\mathbf{b}$, and the distance matching deals with these information.

- Since $\mathbf{b}$ contains information about the singularity center location within its windows $W(sing_s)$, we convert this center in a pixel position (x, y).
- Since $\mathbf{A}$ contains information about the singularity type (fig 1), matching the coefficients of $\mathbf{A}$ leads to match singularities of the same type or at least with close vector field aspects.

The entire algorithm to extract singularities and match them as a singlet is described in algorithm 2. An example of singlet is presented in figure 3.

Collecting singlets along a video and analyzing their shapes, positions or their lengths provide robust description of optical flow within this time lapse. Besides, in the next section, we detail our application of singlets description on soccer videos in order to sum up sport matches.

foreach *frame in the shot* **do**
  get the frame f_n;
  if *previous frame exists* **then**
    compute the optical $flow_n$=(U,V) from f_{n-1} and f_n;
    foreach *sliding window* **do**
      crop the optical flow within the window;
      compute the projection of each component U and V by doing the scalar product of each component with each element $P_{i,j}$ of the Legendre basis;
      change of basis to get $\mathbf{A}$ and $\mathbf{b}$;
      detect if there is a singularity and its characteristics;
    end
  end
  foreach *detected singularity* $sing_s$ *in* $flow_n$ **do**
    foreach *singularity of* $flow_{n-1}$ *within* $V(sing_s)$ **do**
      find the one with the best distance between their $\mathbf{A}$s and the singularities positions;
    end
    if *the bestdistance is under 2* **then**
      if *the bestMatch is at the end of a singlets* **then** add $sing_s$ in this singlets;
      else create a singlets with the both singularities.;
      if *no match has been found within* $flow_n$ **then**
        search a match in the previous flow until an historic of 5 flows
      end
    end
  end
end

Algorithm 2: Algorithm to extract singlets

4. Experiments

4.1. Data

Facing the lack of benchmark from TV in the domain of sport video analysis, we have recorded our own videos from HDTV broadcasting. We manually annotated zooms and slow motions from 4 soccer matches of the FIFA World CUP 2014 (Germany vs Portugal, Nigeria vs Argentine, France vs Honduras, Switzerland vs France) and the Qatar Handball World Cup 2015 final. Each video has been scaled to 25 fps.

In order to evaluate our salient moment detection, we have extracted the ground-truth for the 4 selected matches (Germany vs Portugal, Nigeria vs Argentine, France vs Honduras, Switzerland vs France) from lists of salient moments extracted on the official FIFA website [3]. As can be seen from these ground truths, the description of each salient event is linked to a given minute of the match and describes roughly the action. We manually annotated the salient moments from the FIFA description with the number of starting and ending frames to obtain a frame level decision.

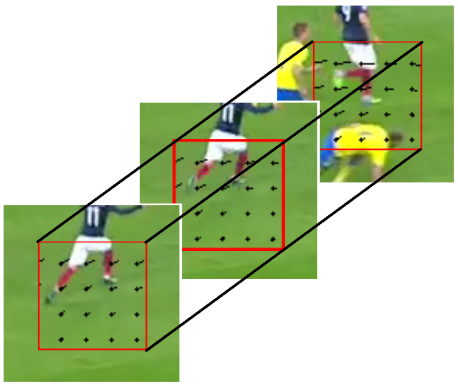


Figure 3. Singlets: illustration of matching of singularities extracted from real optical flow on three consecutive frames in a soccer match. It is a spiral singularity (red, as in figure 1) as can be seen from the flow.

Match	FIFA ground-truth	Extended ground-truth
Germany vs Portugal	30	27
Nigeria vs Argentine	51	35
France vs Honduras	54	32
Switzerland vs France	40	26

Table 1. Number of salient moments in each match according to the ground-truth of all the moments described on FIFA official webpage vs our extended ground-truth.

We improved the annotation of this benchmark to make it compliant with a computer vision groundwork. First, we have completed this list of salient moments by adding all corners and kick-off to fill missing actions. Secondly, we have merged the salient moments which are listed twice: for instance, a goal from a kick-off listed as first an action of kick-off and an action of goal. We call the ground-truth with this completed list of salient moments: Extended ground-truth. In table 1, we report the amount of salient moment in each games and in each database.

In this database, there are more than 7 hours, precisely 696002 frames. For reproducible research sake, all our metadata and our code will be available on our website[1].

4.2. Zoom detection

In a soccer match from TV channel, zooms, and in particular zoom(s)-in(s), are effective indicators of highlights and salient moments of the match, since they represent a

[1] http://www.i3s.unice.fr/~kblanc/

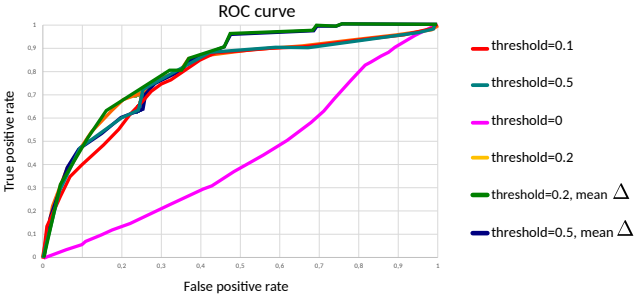


Figure 4. ROC curves of our method for different thresholds on $\det(\mathbf{A})$. Best result is obtained when there is a threshold of 0.2 on $\det(\mathbf{A})$ and an averaged $\Delta(\mathbf{A})$ on 30 frames.

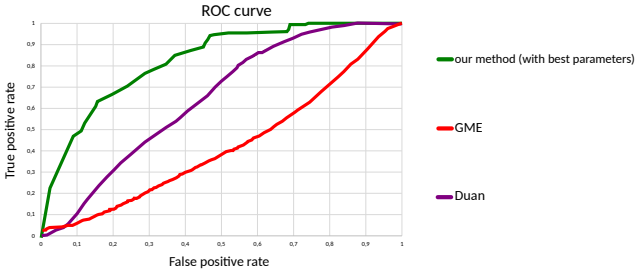


Figure 5. ROC curves for each methos for zoom detection: our method vs Duan method vs GME method.

Method	Precision	Recall	Accuracy
GME	3.68 %	68.4 %	19.79 %
Duan	8.92 %	50.62 %	75.06 %
ours	19.45 %	63.47 %	86.82 %

Table 2. Precision, recall and accuracy for zoom detection.

natural reaction of the cameraman to an action salience.

To evaluate, we use a video of the half soccer game, *i.e.* more than 128500 frames. We only annotated persistent zooms which are important in intensity and in time. There are exactly 5659 positive optical flows and 122841 negative optical flows. To compare each methods, we use ROC curves (figure 4 and 5) which well illustrate performances, even in the case of unbalanced classes.

The two singularities *star node* and *improper node* (figure 1) represent zooms. Two conditions are required for these detection: a strong determinant $\det(\mathbf{A})$ and a low $\Delta(\mathbf{A})$ (cf eq. 3). det(A) corresponds to the intensity of the optical flow. In figure 4, we evaluate the influence of a threshold on $\det(\mathbf{A})$ over zoom detection performance: no threshold on $\det(\mathbf{A})$ implies that every global motion is a zoom candidate, even slight ones and as one could guess, that option gives poor results; with a threshold on $\det(\mathbf{A})$, results are quite stable. Since projections are all computed from optical flows and in order to get a detection with time consistency closer to human perception, we add a chronological window to average these $\Delta(\mathbf{A})$ during time. After grid search on the window size from 5 to 100, best results are obtained with a time history of 30 frames, that almost corresponds to a second. We obtain best results with a chronological average, a threshold of 0.2 on $\det(\mathbf{A})$ and a threshold of 4 on $\Delta(\mathbf{A})$ (figure 4).

Usually zooming are detected by the Global Motion Estimation (GME) method [30]. For this method, we used the RGMC [20] method to compute the homography h_t that models the camera motion for each frame t.

$$x' = m_0 x + m_1 y + m_2$$
$$y' = m_3 x + m_4 y + m_5$$

According to [16], a zoom is then detected if m_0 and m_4 are equals. We set then a threshold on their difference. The best threshold value found for GME method is 0.0004.

In [5], Duan quantified motion vectors to produce two histograms, on angles and on magnitudes. A zoom is detected if there is enough vectors with small norms and enough vector angles between the following angles $[15; 75] \cup [105; 165] \cup [195; 255] \cup [285; 345]$. Therefore, this method requires two thresholds which are not provided. The detection gives best results without filtering the magnitude histogram (i.e. no threshold on the magnitude histogram) leading us to focus on the angle histogram. The best threshold value for Duan method on the angle histogram is 42% of the pixels numbers.

We compare our method with the best parameter setting with reference approaches but we can see by looking at figures 4 and 5 jointly that our method is not very sensitive to the parameter setting since most of the settings provide better results than reference approaches.

In the table 2, we compute the accuracy of each method with the best corresponding threshold. Since our database is unbalanced with 5% of positives samples, the precision values are weak but are still a good indicator of recognition rate to check if the classifier does not always answer negative classes.

Our zoom detection method has three advantages. The main advantage of our method is that zooms are detected even if the zoom direction is not in the image center. As can be seen in figure 6, the optical flow center (represented as the light blue dot on the third row) is very far from the zoom center which is still detected. To the best of our knowledge, such a result can only be achieved with our approach. Moreover, we can also localize the zoom center which is an indicator of where the action happens. Finally, the last advantage is to easily differentiate zoom-in and zoom-out. In the case of those singularities, *star node* and *improper node*, the eigenvalues of $\mathbf{A}$ are equals. Thus we simply have to check the sign of one of the eigenvalues to differentiate zoom-in and zoom-out. This method is very efficient to detect zoom anywhere in the video.

Therefore, extracting global singularity on the whole op-

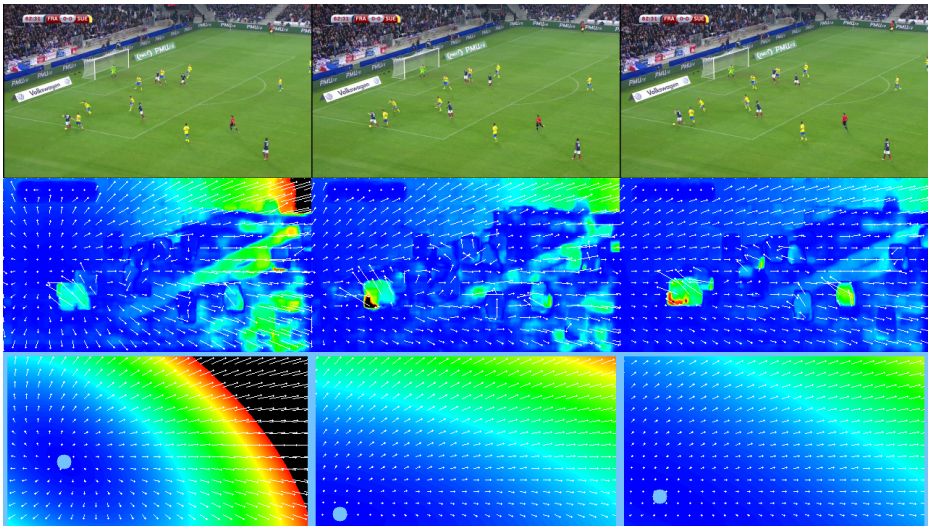


Figure 6. A zoom detected in a soccer match. From left to right: 3 consecutive frames. From top to bottom: original frames (ball highlighted by a red square), original optical flows and polynomial optical flow approximations. In the last rows, the singularity center is represented by a small target and it is centered on the soccer ball that you can see in the first row.

foreach *frame in the shot* **do**
  get the frame f_n;
  if *previous frame exists* **then**
    compute the optical flow$_n$=(U,V) from f_{n-1} and f_n;
    compute the projection of each component U and V by doing the scalar product of each component with each element $P_{i,j}$ of the Legendre basis;
    change of basis to get $\mathbf{A}$ and $\mathbf{b}$;
    if $|\det(\mathbf{A})| > 0.2$ *and* $|\Delta(\mathbf{A})| < 4$ **then**
      zoom detected
      zoom center position at $-\mathbf{A}^{-1}\mathbf{b}$ **if** $trace(\mathbf{A}) < 0$ **then** zoom-out detected;
      else zoom-in detected;
    end
  end
end

Algorithm 3: Zoom detection

tical flow can provide useful video editing information such as zoom detection while extracting local singularities help to measure the amount of motion.

4.3. Global excitement

We made the hypothesis that a salient moment of team sports is when several players are in conflict to get the ball or when players are running towards the goal (attack). In that cases, there are local singularities around moving zones and these singularities will have a temporal consistency.

We detect singularities in each optical flow within a sliding window and compute a spatial histogram on their positions. Our purpose is to detect regions where there is an global excitement. Each image is splitted by a 3x3 histogram, 9 uniform bins (see figure 7).

In order to stabilize spatial histograms through time, we sum spatial histograms within a temporal window of size 10 frames. We discard video regions where there is a scoreboard and sum up the remaining bins. We choose this indicator to select moments for a summary and sort them according to their intensity based on the aforementioned hypothesis. These histograms are particularly interesting when they are extracted from the farthest view in a match.

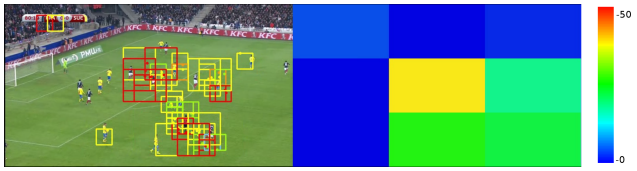


Figure 7. Left: a shot with all detected singularities. Each colored square corresponds to one singularity, with the color of their type (figure 1). Right: their corresponding space-time histogram expressed with the given heat scale representing the amount of singularity.

Other clues indicate that an important action just happened such as replays (slow motions).

4.4. Slow motion detection

When a fast action deserves to be detailed, producers usually use slow motions, often from an other point of view

Dataset	Precision	Recall	Accuracy
Train	97.06 %	80.49 %	89.41 %
Test	76.32 %	87.88 %	79.36 %
Test on handball	100 %	20 %	60 %

Table 3. Precision, recall and accuracy for slow motion detection on different datasets.

to see all details. Thus we can detect the match salient moments in spotting slow replay. Most of the proposed solutions for this detection problem use the video production habits that consist usually in putting a logo before and after each replay. Pan *et al.* [15] used a HMM algorithm to detect slow motion from differences between frames and an HMM to model states before and after the slow motion like editing effects for shot transition. Zawbaa *et al.* [32] filter each frame of the video with the logo dominant color and a trained SVM predicts if the logo appears or not. This type of method has the drawback of both requiring a training set of logo appearance and totally ignoring that the motion is very slow.

To detect slow motion, we focus on the slowness intensity of motion. Our method is based on a simple assumption that is: in a slow replay, a motion will be consistent during time. A singlet describes a typical motion evolution until it totally change its type. Therefore the length of singlets tells us how long a motion last. Thus several long singlets implies that the shot is a slow motion.

We compute the length of each singlet and store their length in a histogram of 100 bins. The histogram is then considered as the feature vector for detecting video slowness. To detect slow motion in video, we train a SVM with a radial basis kernel function after whitening and a PCA on our slow motion feature vector. To train an SVM, we use a database of 82 videos containing 41 slow motion and 41 non slow motion video parts extracted from three of the soccer matches. In our experiment, we set the parameters as $\alpha = 0.4$ for the neighborhood selection (cf eq 4), $\lambda = 0.02$ for the matching distance (cf eq 5) and $C = 29$ and $\gamma = 4.10^3$ for the SVM training.

The test sets is composed by the rest of slow motion and non slow motion samples in each soccer matches, being 33 slow motion segments and 33 non slow motion segments. In order to evaluate the power of generalization of our descriptor, we add 5 slow motions and 5 non slow motions part extracted from the handball game.

Table 3 presents our recognition results on each slow motions dataset. As you can see, we obtain high accuracy for soccer slow motions classification. With the same feature extraction framework and without retraining the SVM, we obtain a slow motion detection at high precision score for the handball dataset. Since our framework was trained on soccer matches and since we do not change any parameters, regarding the results on the handball game, it is indeed generic in spite of an over detection effect.

In the next part, we describe how singlets are useful through these detections to extract salient moments of a match and make a summary.

4.5. Salient moment detection and match summarization

Singularities and singlets represents motion regions. Globaly extracted, singularity represents camera motion: we use star node and improper node to spot zooms in section 4.2. Locally extracted, they represents players and ball mouvements: we use the amount of singularities to characterize global excitement. Singlet identifies singularity's evolution during time: we use their length to notice slow motions by their nature.

From the singlets' descriptions, we compute a summarization. We select best moments by a combination of several zoom-in and zoom-out, followed by a peak in the space time histogram of singularities and then a replay of the action with a slow motion.

In order to combine different clues of the salient moments, we build a frieze for each different values. In figure 8, you can see on the left a soccer video frame and on the right its actual singlets space-time histogram. Underneath, from top to bottom, there are 5 friezes. The first frieze corresponds to extracted zooms: red for zoom-in and blue for zoom out. The second frieze shows the quantity of zoom changes within a second according a heat scale. The next frieze shows the amount of large vectors also using a heat scale. We use this light indicator to contrast close-up view from large field view. The forth frieze indicates SVM classification for slow motions in close up views: red for slow motion and blue for non slow motion. Finally the last frieze indicates the amount of singularities: the global excitement indicator.

A **match summarization** is the concatenation of detected salient moments. A salient moment is detected if within 30 seconds time frame there are:

- at least two zoom direction changes, and
- an activity peaks higher than 1500 (at least 1500 singularities) in a farthest view, and
- a slow motion replay in a close up view.

Therefore, we extract a list of main moments of the match with these rules. On these 4 matches, our method obtains the results referred in table 4.

Unfortunately, the authors of summarization methods for soccer games do not provide any source code nor any binary executable to reproduce their results on our database and the re-implementation implies either human user specifications,

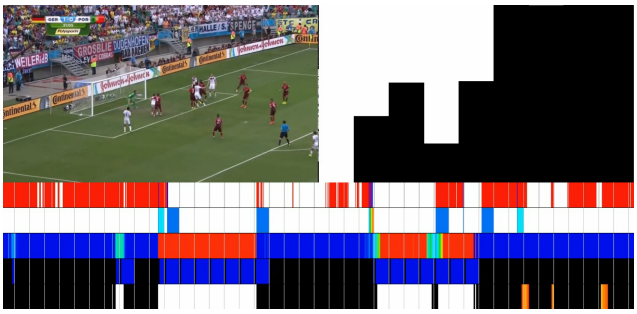


Figure 8. Top left: random frame of a soccer match. Top right: its space time histogram. Bottom: 5 friezes that describe a salient moment detection criteria (from top to bottom: zoom-in/out, zoom changes count, saturation, slow motion detection and activity score).

Match	FIFA ground-truth	Extended ground-truth
Germany vs Portugal	80 %	88.9 %
Nigeria vs Argentine	53 %	77.2 %
France vs Honduras	53.7 %	90.7 %
Switzerland vs France	62.5 %	96.6 %
Mean	62.3 %	88.2 %

Table 4. Precision rate of detected salient moments by our method over all the salient moments, using two different ground-truth (see paragraph 4.1).

not provided thresholds or logo database collection. This makes the comparison impossible for video summarization.

We can summarize a soccer match by extracting zooms, by detecting slow motion and salient moments without any hypothesis on the player movements, just in analyzing global video motion. To compute the summary, the shots that contains the detected activity peaks are aggregated.

Since we do not use either soccer characteristics or producing specificities (e.g. logos), our approach is generic. In order to confirm this, we have set all parameters of our method for soccer and train the SVM on soccer video training set. We have then extracted salient moments on a extract of an handball match without any retraining, refining or adaptation. The extract is a part the HDTV video of the final of the 2015 world championship, Qatar against France (figure 9). We detect on that extract one salient moment, followed by one slower replay, out of the three which are indeed in this extract ground-truth. Please notice that the handball field is purple, and that the players are moving differently with zone restriction rules really different from soccer games.

5. Discussion

In this article, we focus on detecting salient moments as zooms, slow motion replays and global excitement of players. However our description gives other information within the projection coefficients that could lead to a semantic description of a game.



Figure 9. Frame of the extract of handball game.

For instance, the simple degree 0 on polynomial basis (cf eq.1) can provide interesting information in the process of semantic description: $u_{0,0}$ and $v_{0,0}$ gives the translation on the vertical and the horizontal axis. These coefficients from the projection of the global optical flow, as computed for zoom detection 4.2, characterize the camera translation and then they determine the presence of an attack phase or a counterattack phase.

These projections can be done from any bivariable function to any degree of polynomial bivariable function. We are currently studying these coefficients to provide other clues on a soccer match.

In the open code, we provide the possibility to set the degree of polynomial projection, so these other motion features can be easily computed.

6. Conclusion

We have presented in this article a new robust multi-scale video descriptor: the singlets. They correspond to the tracking of singularities in the polynomial projections of optical flow along the temporal dimension of the video.

This descriptor proves its informativeness in detecting zooms (in and out), slow motions and salient moments during sport events without any ad-hoc elements (no logo, no particular ground color), thus allowing to build relevant sport summary.

On future works, we will evaluate the potential of singlets for a retrieval task as a motion descriptor itself or as a complementary description to a shape and color description.

Acknowledgment

We are grateful to Wildmoka for fruitful discussions

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016.

[2] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati. Semantic annotation of soccer videos: automatic highlights identification. *CVIU*, 92(2):285–305, 2003.

[3] Fédération Internationale. de Football Association (FIFA). 2014 fifa world cup brazil: matches description. Germany-vs-Portugal: `www.fifa.com/worldcup/matches/round=255931/match=300186475/live-blog.html` Nigeria vs Argentine `www.fifa.com/worldcup/matches/round=255931/match=300186458/live-blog.html` France vs Honduras: `www.fifa.com/worldcup/matches/round=255931/match=300186496/live-blog.html` Switzerland vs France: `www.fifa.com/worldcup/matches/round=255931/match=300186514/live-blog.html`).

[4] L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian, and C.-S. Xu. A mid-level representation framework for semantic sports video analysis. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 33–44. ACM, 2003.

[5] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin. A unified framework for semantic shot classification in sports video. *IEEE Transactions on Multimedia*, 7(6):1066–1083, 2005.

[6] A. Ekin. Generic play-break event detection for summarization and hierarchical sports video analysis. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 1, pages I–169. IEEE, 2003.

[7] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, pages 363–370. Springer, 2003.

[8] Y. Gong, L. T. Sin, C. H. Chuan, H. Zhang, and M. Sakauchi. Automatic parsing of tv soccer programs. In *Multimedia Computing and Systems, 1995., Proceedings of the International Conference on*, pages 167–174. IEEE, 1995.

[9] A. Hanjalic. Generic approach to highlights extraction from a sport video. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1, pages I–1. IEEE, 2003.

[10] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):797–819, Nov 2011.

[11] O. Kihl, B. Tremblais, and B. Augereau. Multivariate orthogonal polynomials to extract singular points. In *IEEE International Conference on Image Processing 2008. ICIP 2008.*, pages –, San Diego, CA, United States, Oct. 2008.

[12] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, Sept. 2005.

[13] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 39(5):489–504, 2009.

[14] R. Leonardi, P. Migliorati, and M. Prandini. Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled markov chains. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):634–643, 2004.

[15] H. Pan, P. Van Beek, and M. I. Sezan. Detection of slow-motion replay segments in sports video for highlights generation. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 3, pages 1649–1652. IEEE, 2001.

[16] X. Qian. *Global Motion Estimation and Its Applications*. INTECH Open Access Publisher, 2012.

[17] M. Raptis and S. Soatto. *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I*, chapter Tracklet Descriptors for Action Modeling and Video Analysis, pages 577–590. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[18] A. Raventos, R. Quijada, L. Torres, and F. Tarres. Automatic summarization of soccer highlights using audio-visual descriptors. *arXiv preprint arXiv:1411.6496*, 2014.

[19] D. Sadlier, N. E. O'Connor, et al. Event detection in field sports video using audio-visual features and a support vector machine. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(10):1225–1233, 2005.

[20] S. M. Safdarnejad, X. Liu, and L. Udpa. Robust global motion compensation in presence of predominant foreground. In *BMVC*, pages 21–1, 2015.

[21] Y. Tabii and R. O. Thami. A new method for soccer video summarizing based on shot detection, classification and finite state machine. In *Proceedings of The 5th international conference SETIT*, 2009.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[23] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(1):3, 2007.

[24] V. Vijayakumar and R. Nedunchezhian. A study on video data mining. *International journal of multimedia information retrieval*, 1(3):153–172, 2012.

[25] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

[26] C. Wu, Y.-F. Ma, H.-J. Zhan, and Y.-Z. Zhong. Events recognition by semantic inference for sports video. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 805–808. IEEE, 2002.

[27] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun. Structure analysis of soccer video with hidden markov models. In

Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, volume 4, pages IV–4096. IEEE, 2002.

[28] Z. Xiong, X. S. Zhou, Q. Tian, Y. Rui, and T. S. Huang. Semantic retrieval of video. *IEEE Signal Processing Magazine*, 23(2):18, 2006.

[29] P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, and H. Sun. Algorithms and system for segmentation and structure analysis in soccer video. In *ICME*, volume 1, pages 928–931, 2001.

[30] Q. Ye, Q. Huang, W. Gao, and S. Jiang. Exciting event detection in broadcast soccer video with mid-level description and incremental learning. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 455–458. ACM, 2005.

[31] D. Yow, B.-L. Yeo, M. Yeung, and B. Liu. Analysis and presentation of soccer highlights from digital video. In *proc. ACCV*, volume 95, pages 499–503, 1995.

[32] H. M. Zawbaa, N. El-Bendary, A. E. Hassanien, and T. Kim. Event detection based approach for soccer video summarization using machine learning. *International Journal of Multimedia and Ubiquitous Engineering*, 7(2):63–80, 2012.