Automatic Detection and Recognition of Trademarks in Sports Videos

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, and Arjun Jain

Media Integration and Communication Center, University of Florence, Italy [a]

[a]This work is partially supported by Sport System Europe srl, Bologna (Italy), www.sportsystem.com

1. Introduction

***Goal:* Detect, recognize and perform robust localization of trademarks in sports videos.**

- Traditional trademark recognition systems deals with the problem of content-based indexing and retrieval in logo databases with the goal of assisting the process of trademark registration.
- In this case the image acquisition and processing chain is controlled so that the images are of acceptable quality and are not distorted.
- The problem of trademark recognition in real world videos is inherently harder, due to the relatively low quality of the images (e.g. video interlacing, color sub-sampling, motion blur, compression artifacts, etc.).
- The appearance of trademarks in sports videos are often characterized by **perspective deformations**, **motion blur** and **occlusions**.

2. The System

We propose a semi-automatic system for detecting and retrieving trademark appearances in sports videos. A human annotator supervises the results of the automatic annotation through an interface that shows the time and the position of the detected trademarks; due to this fact the aim of the system is to **provide a good recall figure**, so that the supervisor can safely skip the parts of the video that have been marked as not containing a trademark, thus speeding up his work.

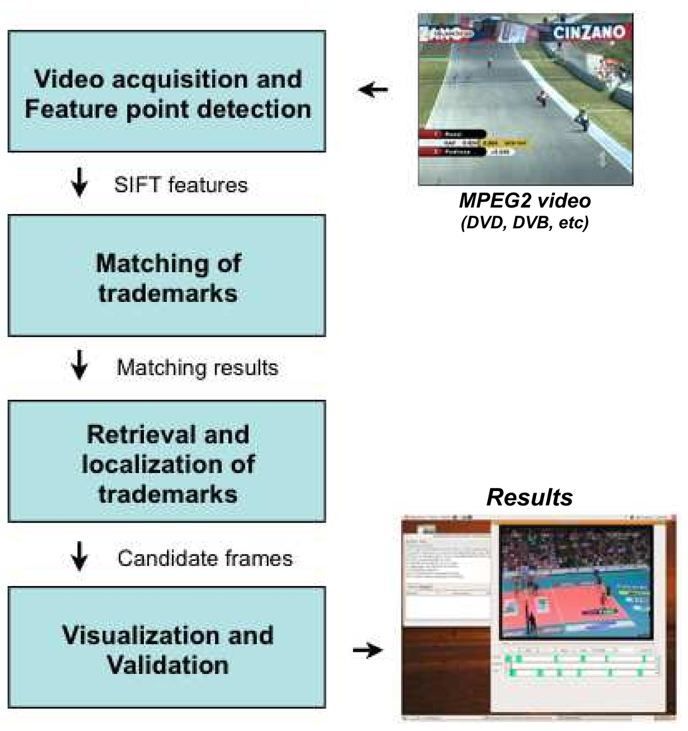

Figure 1: *Overview of our system.*

3. Approach

The first challenge is to **build a model that is able to cope with partial occlusions**.

- Therefore, we use DoG points and SIFT feature descriptors as a compact representation of the important aspects and local texture in trademarks.
- By combining the results of local point-based matching we are able to match entire trademarks.
- A supervised machine-learning approach is used to dynamically adapt the similarity threshold used to assess the trademark matches.

The process of interest points detection and description is time consuming.

- MPEG2 videos (25*fps*) are sub-sampled and SIFT points are detected at 5*fps*; these frames are selected measuring their visual quality.
- Visual quality is estimated measuring blurriness, number of edges and number of SIFT points (these are used as a hint to evaluate the likelihood of the detectability of trademarks).
- Frame classication (selection) has been performed by SVM using a RBF kernel.

4. Image and video features

Trademarks are represented as a **bag of SIFT feature points** and each trademark is represented by one or more graphical instances.

Trademark T_j is so represented by the N_j SIFT feature points detected in the image:

$$T_j = \{(x_k^t, y_k^t, s_k^t, d_k^t, \mathbf{O}_k^t)\}, \text{ for } k \in \{1, \ldots, N_j\},$$

and x_k^t, y_k^t, s_k^t, and d_k^t are, respectively, the x- and y-position, the scale, and the dominant direction of the kth detected feature point; O_k^t is a 128-dimensional local orientation histogram of the SIFT point (t is used only to distinguish points from trademarks and video frames).

Each frame, V_i, of a video is represented similarly as a bag of M_i SIFT-feature points detected in frame i.

5. Detection and retrieval of trademarks

Detection and retrieval of trademarks is done by comparing the bag of local features representing the trademark T_j with the local features detected in the frames of the video V_i.

For each point in T_j we search for its two nearest neighbors N_1 and N_2 in the V_i point set:

$$\begin{aligned} N_1(T_j^k, V_i) &= \min_q ||\mathbf{O}_q^v - \mathbf{O}_k^t|| \\ N_2(T_j^k, V_i) &= \min_{q \neq N_1(T_j^k, V_i)} ||\mathbf{O}_q^v - \mathbf{O}_k^t||. \end{aligned}$$

and we compute its *match score*:

$$M(T_j^k, V_i) = N_1(T_j^k, V_i)/N_2(T_j^k, V_i)$$

The *match set* for trademark T_j in frame V_i is so defined as: $M_i^j = \{k \,|\, M(T_j^k, V_i) < 0.8\}$.

The final determination of whether a frame V_i contains trademark T_j is made by thresholding the *normalized match score*:

$$|M_i^j|/|T_j| > \tau \iff \text{trademark } T_j \text{ present in frame } V_i$$

The **normalized match threshold (NMT)** τ requires that a certain percentage of the feature points detected in the reference trademark T_j must be matched to the frame V_i.

- A value of ~ 0.2 is a reasonable choice for several different sports (Fig. 2).
- Analysis of the precision–recall curves allows to determine the best choice for *NMT* (e.g. Fig. 4).

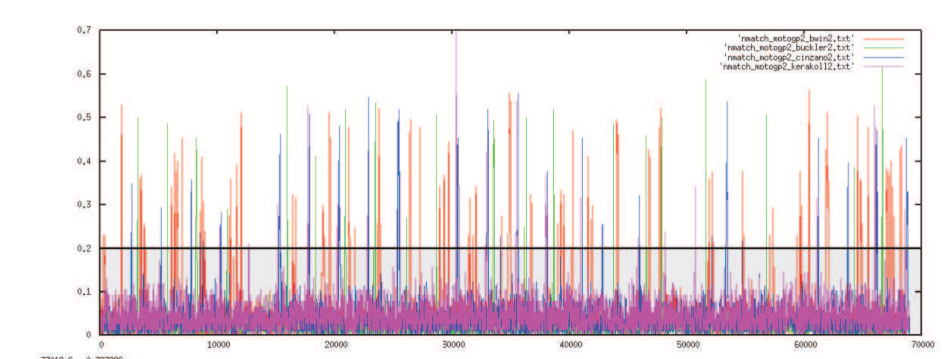

Figure 2: *Normalized match score histograms of 4 different trademarks in a MotoGP video.*

6. Robust Trademark Localization

In order to localize the trademark in the original frame V_i and to approximate its area, we compute a robust estimate of the feature point cloud (Fig. 3).

The current feature point locations are so denoted as $F = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

The robust centroid estimate is computed by iteratively solving for (μ_x, μ_y) in

$$\sum_{i=1}^{n} \psi(x_i; \mu_x) = 0, \quad \sum_{i=1}^{n} \psi(y_i; \mu_y) = 0$$

where the influence function ψ used is the Tukey biweight:

$$\psi(x; m) = \begin{cases} (x-m)(1 - \frac{(x-m)^2}{c^2})^2 & \text{if } |(x-m)| < c \\ 0 & \text{otherwise} \end{cases}$$

The scale parameter c is estimated using the *median absolute deviation from the median*: $\mathrm{MAD}_x = \mathrm{median}_i(|x_i - \mathrm{median}_j(x_j)|)$.

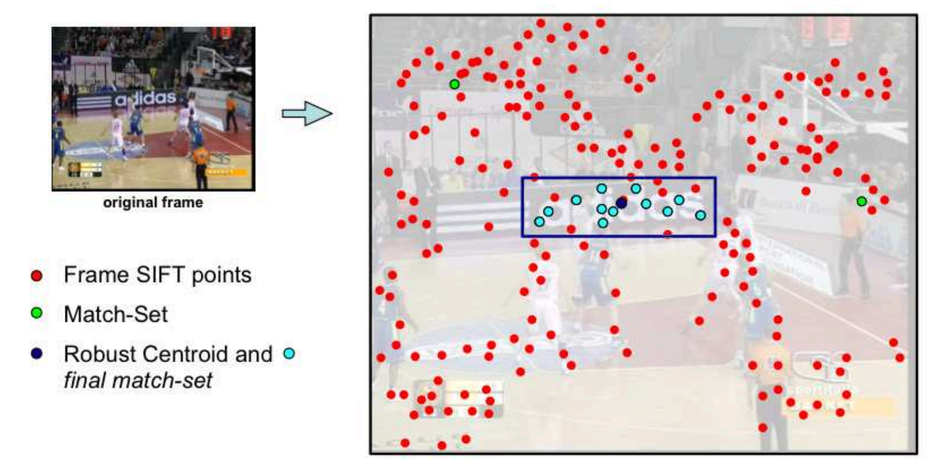

Figure 3: *Robust trademark localization*

7. Automatic *NM Threshold* adaptation

The initial (static) value of the *normalized match threshold* τ is dynamically adapted to take into account the frame visual quality.

- We have performed experiments to determine what is the lowest acceptable value τ_{min} (0.08 in our experiments) for the *NMT*.
- An SVM classifier has been trained to automatically select the value of the *NMT* that gives the best trade-off between precision and recall.

8. Results

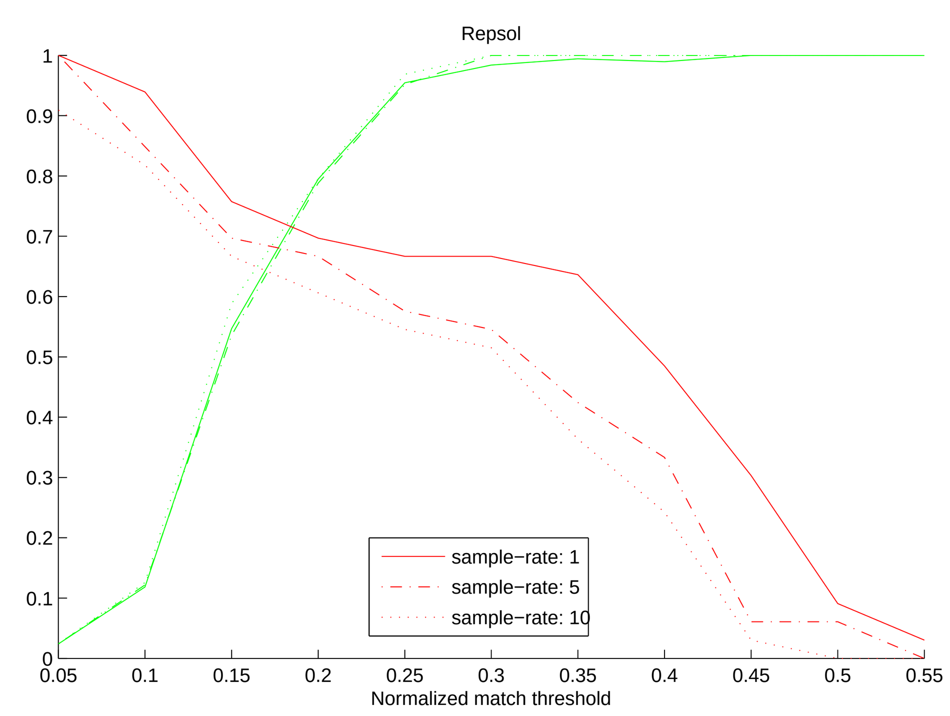

Figure 4: *Precision (green) and recall (red) as a function of the NMT and of the sample-rate.*

- Experiments have been performed on several videos of different sports (motogp, formula-one, volleyball, soccer and basket); in most cases a precision rate of about 85% can be achieved with a recall of around 60%.
- However, the recall figure show great variations (in particular it happens in sports like Volleyball and Basket).
- The use of automatic *NMT* adaptation provided by a SVM classifier improves the recall of our system ($\sim +15\%$) with a minimal cost in terms of precision ($\sim -5\%$).