# VISUAL ANALYSIS FOR DRUM SEQUENCE TRANSCRIPTION

*Kevin McGuinness†, Olivier Gillet‡, Noel E. O'Connor†, Gaël Richard‡*

† Centre for Digital Video Processing, Dublin City University, Glasnevin, Dublin 9, Ireland
‡ GET / Télécom Paris (ENST). CNRS LTCI. 37 rue Dareau, 75014 Paris, France

## ABSTRACT

*A system is presented for analysing drum performance video sequences. A novel ellipse detection algorithm is introduced that automatically locates drum tops. This algorithm fits ellipses to edge clusters, and ranks them according to various fitness criteria. A background/foreground segmentation method is then used to extract the silhouette of the drummer and drum sticks. Coupled with a motion intensity feature, this allows for the detection of 'hits' in each of the extracted regions. In order to obtain a transcription of the performance, each of these regions is automatically labeled with the corresponding instrument class. A partial audio transcription and color cues are used to measure the compatibility between a region and its label, the Kuhn-Munkres algorithm is then employed to find the optimal labeling. Experimental results demonstrate the ability of visual analysis to enhance the performance of an audio drum transcription system.*

## 1. INTRODUCTION

Music transcription is the process by which a high-level description of a music piece, typically a music score, is extracted from an audio signal. Transcription plays a fundamental role in content-based music retrieval systems, but also finds application in computer-aided music teaching or composition. Transcription is a challenging task and has been the subject of significant research – see [1] for a comprehensive review. An interesting subproblem is *polyphonic drum transcription* (or *drum event detection*), which consists of extracting from complex polyphonic music recordings the times at which each instrument of the drum kit (bass drum, snare drum, tom) is played.

In this paper, we investigate the use of video information for the purpose of drum transcription. Video analysis systems have been proposed for a variety of music related tasks, such as baton tracking [2], expression analysis in clarinet playing [3], or accent analysis in drumming [4]. Multimodal systems specifically tailored for audiovisual transcription tasks are described in [5] for piano performances or [6] for drums. In this paper, we present several key improvements of this prior work. Since [6] employs a statistical machine learning approach with scene-dependent features, a generic classifier cannot be learned and a specific classifier must be retrained for each sequence processed using either a reference sequence (manually annotated), or a partial audio transcription. Our approach here, however, considers the video analysis task as a detection process rather than as a classification problem thus removing the need for training sequence-dependent classifiers. Furthermore, even though an automatic calibration process is described in [6], the system does not make use of high-level video features – no effort is made to understand the semantics of the image. Thus, we propose improved video analysis with emphasis on the detection of the position of drum instruments within the scene. Finally, we evaluate our work on a more diverse, large and difficult database.

The paper is organized as follows. The next section gives an overview of the video analysis system. The following four sections present each of its components. Section 7 discusses the results obtained and, finally, section 8 suggests some conclusions and future directions.

## 2. OVERVIEW

Several steps (detailed in figure 1) are required to efficiently detect drum hits. Firstly, the video sequence is analyzed to detect the position of each drum element (drums and cymbals) in the scene, and more specifically the part of the instrument hit by the drum sticks (referred to as *drum tops* throughout this paper). Since all the drum tops have a circular shape, the most efficient criterion for this task is geometric. Then, a simple motion intensity feature, coupled with foreground object segmentation is used to detect drum strokes on each of the detected drum tops. In order to obtain a transcription, it is necessary to identify which drum instrument corresponds to each detected drum top. Cymbals and other drums (toms, snare drum) are discriminated using hue features. We also investigate the use of an audio drum transcription system as an additional source of information to unequivocally assign each detected region to the corresponding drum instrument. Finally, once a video transcription is obtained, it can be fused with an audio transcription, or other video transcriptions obtained from different cameras.

## 3. SEGMENTATION OF DRUM TOPS

We employ an ellipse detection algorithm to automatically extract the locations of the drum tops. The algorithm begins by extracting edge pixels and clustering those likely to belong to the same edge in an image. Then, ellipses are fitted to groups of edge clusters resulting in a potential ellipse. The hypothesised ellipse is then examined, and several fitness metrics are computed to determine if the prospective ellipse corresponds to a real ellipse. If we determine that an ellipse is valid according to the above criteria, we remove the clusters used to fit the ellipse and update the fitness functions to ignore pixels that lie within the found ellipse. The searching of edge clusters is then restarted, allowing for the detection of ellipses which were previously occluded by the presently detected ellipse.

### 3.1 Edge Detection

To extract potential edge pixels, we employ a modified version of the well known Canny operator [7]. Our modified canny operator works in much the same way as the original, with two important differences. First, we replace the standard Gaussian smoothing operation with a Gaussian Bilateral Filtering [8]. This helps to alleviate noise while preserving significant image discontinuities. Second, our algorithm determines the magnitude of the discontinuities as the weighted average of $L*u*v$ color differences in a $3\times 3$ window, instead of the usual grayscale Sobel operator. Once the edge map $E(x,y)$ has been computed, we then threshold with a high and a low threshold, producing a coarse edge map $E_c(x,y)$ and a fine edge map $E_f(x,y)$.

### 3.2 Edge Clustering

Using the coarse edge map $E_c(x,y)$, we wish to produce a set of pixel clusters $C=\{C_0,\ldots,C_n\}$, such that each cluster contains pixels that are likely to belong to the same semantic edge in the image. The first step is to cluster together all connected pixels using the 8-connectivity constraint. The result of this operation may, of course, produce clusters of edge points that form edges of multiple objects, due to noise and occlusions so we further segment these clusters at points of local curvature maxima. To achieve this, we

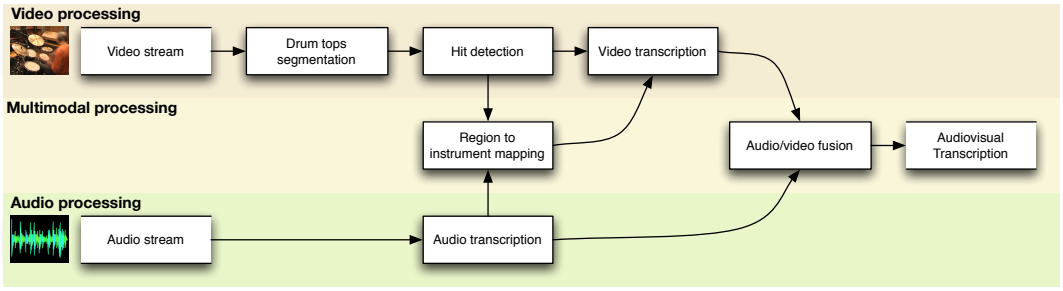


Figure 1: Overview of the audio/video analysis

first estimate curvature for each point in a cluster. However, direct estimation curvature using first and second derivatives of the planar parametric curves in $C_i$ produces quite inaccurate results, due to noise and quantization error. Observing that curvature for a plane curve at a given point has a magnitude equal to the reciprocal of the radius of an osculating circle; we can estimate curvature at each discrete point along the curve by estimating the osculating circle at that point. The osculating circle at a given point can be approximated by fitting a circle to points in the same cluster within a given radius of the point. We use the approximated Euclidean distance circle fitting method described in [9] to fit the circle.

### 3.3 Ellipse Fitting

To select clusters of pixels and approximate the ellipses that they imply, we simply iterate through all possible individual clusters and pairs of clusters, beginning with the largest. However, a randomized selection process, such as RANSAC [10] could also be used at this stage (see [11]). Once some data points are selected, we fit an ellipse to them using the direct least mean square method, proposed in [12]. In brief, for a set of observed points $S = \{\langle x_i, y_i \rangle : 0 < i \leq n\}$ the method finds the set of ellipse parameters $\mathbf{a} = [a\ b\ c\ d\ e\ f]^T$ that minimizes the sum of square algebraic distances, subject to the ellipse constraint. That is, subject to $b^2 < 4ac$ it is the solution to,

$$\arg\min_{\mathbf{a}} \sum_{i=1}^{n} (\mathbf{a}^T \mathbf{x}_i)^2 \tag{1a}$$

$$\mathbf{x}_i = \begin{bmatrix} x_i^2 & x_i y_i & y_i^2 & x_i & y_i & 1 \end{bmatrix}^T \tag{1b}$$

In [12] it was shown that this problem can be solved by expressing it as a generalized eigenvalue problem

$$\mathbf{D}^T \mathbf{D}\mathbf{a} = \lambda \mathbf{C}\mathbf{a} \tag{2}$$

where $\mathbf{D}$ is the *design matrix*, and $\mathbf{C}$ is a $6 \times 6$ *constraint matrix* expressing the ellipse constraint as the equality $4ac - b^2 = 1$, or in quadratic form $\mathbf{a}^T \mathbf{C}\mathbf{a} = 1$.

$$\mathbf{D} = [\mathbf{x}_0 \quad \mathbf{x}_1 \quad \ldots \quad \mathbf{x}_n]^T$$

$$C_{ij} = \begin{cases} 2 & \langle i, j \rangle \in \{\langle 1,3 \rangle, \langle 3,1 \rangle\} \\ -1 & i = j = 2 \\ 0 & \text{otherwise} \end{cases}$$

The eigenvector values corresponding to the single positive eigenvalue $\lambda_i > 0$ in (2) are the ellipse parameters $\mathbf{a}$ that satisfy (1).

### 3.4 Fitness Evaluation

To evaluate the fitness of a potential ellipse, we first need a fast method of approximating which pixels occur within an ellipse, and on its boundary. In our implementation, we use the ellipse rasterization procedure presented in [13] to determine border pixels. To efficiently find the pixels that lie within the ellipse, we use a differential technique similar to [13] to determine the bounding box for the ellipse, and then calculate the distance to the ellipse foci for each pixel within the bounding box. The prospective ellipse is evaluated using four metrics outlined below:

1. *Fit Cost* - We define the fit cost as the average of the Euclidean distances from each point used to fit the ellipse, to the nearest point on the border of the potential ellipse. Let $X = \bigcup C_i$ where $C_i$ is a cluster used to fit the ellipse. If $\mathbf{a}$ are the potential ellipse parameters, the fit cost is defined as,

$$F_c(X) = \frac{1}{|X|} \sum_{x \in X} \text{dist}(x, \mathbf{a})$$

where $\text{dist}(x, \mathbf{a})$ denotes the minimum distance from a point to the ellipse. A large fit cost serves as an indication that the clusters chosen to fit the ellipse are incompatible. To determine the distance from a point to an ellipse, we use the iterative method described in [14].

2. *Discontinuity Fitness* - We define the discontinuity fitness as the average Gaussian Euclidean distance from a point on the ellipse boundary, to an potential edge in the fine edge map $E_f(x,y)$, for a given variance. To efficiently compute the discontinuity fitness, we first compute the square Euclidean distance transform of $E_f(x,y)$ using the linear time method presented in [15]. Then, using this, we compute the Gaussian Euclidean distance transform (GEDT) [16, 17] for each point $x$ in the image as:

$$\text{gedt}_\sigma(x) = \exp\left(\frac{-\|x - x_n\|^2}{2\sigma^2}\right)$$

where $x_n$ is the nearest edge point in the fine edge map, and $\sigma^2$ is the variance. Let $E = \{x_1, \ldots, x_n\}$ be the set of pixels that lie on a rasterized approximation of the ellipse boundary. The discontinuity fitness measure is then defined as:

$$F_d(E) = \frac{1}{n} \sum_{i=1}^{n} \text{gedt}_\sigma(x_i)$$

where $n = |E|$ is the number of pixels that lie on the boundry. Note that the fitness measure $F_d(E) \in (0,1]$ and a fitness of 1 denotes a perfect fit to image discontinities.

3. *Homogeneity* - We define our homogeneity metric as the reciprocal of the color variance for pixels that lie within the ellipse. Homogeneity could also be defined for texture in a more sophisticated implementation. A high degree of color or texture homogeneity within the postulated ellipse serves as a good indicator of fitness.

4. *Application Requirements* - In addition to the specific metrics defined above, a particular application may use a priori knowledge of the features of the ellipses present in the image. In our application, we use a model of drum top colors and count the number of pixels inside the ellipse that match the color model. The model was trained in advance using a C4.5 tree classifier [18].

### 3.5 Handling Occlusions

Once an ellipse has been detected, we can eliminate that clusters contributing to the ellipse from the search, and mark pixels in the interior of the ellipse as "not contributing to cost". The search is then restarted allowing us to detect ellipses in the presence of inter-ellipse occlusions. However, this strategy can lead to spurious detections if we do not determine the degree of occlusion of an ellipse. We compute the degree of occlusion as the ratio of the area of intersection with previously detected ellipses to the area of the current ellipse. Computing the area of intersection of ellipses numerically is non-trivial, so we approximate it using pixel counts within the rasterized ellipses. If the degree of occlusion is large, the detected ellipse is probably spurious.

## 4. DRUM HIT DETECTION

Let $\mathscr{R}_i = \{x_1 \ldots x_{m_i}\}$ be the set of pixels that lie in the interior of the $i$th detected ellipse ($i \in [1, N_R]$). In order to detect drum hits, two visual clues are considered: the impact of a drum stick on the top of the instrument, and the motion of the instrument itself once it has been hit.

### 4.1 Drum stick impact feature

Our initial attempts to track the drum sticks proved to be unsuccessful for several reasons. Firstly, because our test material included sequences played with sticks, mallets, brushes or bundled sticks, no generic color model could be used to segment the sticks from the foregrounds. Secondly, because of the limited frame rate of standard video cameras (our test material was recorded at 25 frames per second), there is significant motion blur between frames preventing the use of tracking methods. Consequently, we used a lower-level approach based on foreground/background segmentation.

The background of the image is modeled as a mixture of 3 gaussians. It is therefore possible to compute for each pixel $x$ in each image frame $n$ the probability $p_B(x,n)$ that this pixel belongs to the foreground. As described in [19], this model can be incrementally adapted. This method efficiently extracts the silhouette of the drummer forearms and the drum sticks. A first set of features measure how many foreground pixels are present in each region of interest, and uses the probability $p_B(x,n)$ as a soft membership function:

$$S_i(n) = \sum_{i=1}^{m_i} p_B(x_i, n)$$

### 4.2 Instrument motion feature

In order to estimate motion intensity, a 5-taps differentiator filter [20] is applied to each pixel brightness sequence $I(x,n)$, to yield a difference image $D(x,n)$. The difference image is thresholded as :

$$D'(x,n) = \begin{cases} |D(x,n)| & \text{if } |D(x,n)| > \tau \\ 0 & \text{otherwise} \end{cases}$$

A motion intensity feature in each region of interest is then defined as:

$$M_i(n) = \sum_{i=1}^{m_i} D'(x_i, n)$$

### 4.3 Detection process

Both $S_i(n)$ and $M_i(n)$ as defined above exhibit strong peaks when the instrument associated to the region of interest $i$ is hit. In the case of $S_i(n)$, the detection function will typically exhibit a narrow triangle-shaped peak, maximal at the time of the impact (as the sticks enters, then leaves the region). In the case of $M_i(n)$ the detection function will typically show a triangle-shaped contribution, along with a decaying component after the time of the impact. The decay of this component is longer in the case of the crash cymbals - as they have more freedom of motion. Let $r_{M_i}(n)$ and $r_{S_i}(n)$ be functions, maximal at $m = 0$, representing these typical peak shapes. We propose the following model for the detection function $S_i$ (the same developments can be applied to $M_i$):

$$\begin{cases} S_i(n) = w(n) & \text{when the instrument is not hit} \\ S_i(n) = w(n) + A r_{S_i}(n) & \text{when the instrument is hit} \end{cases}$$

where $w(n)$ is a gaussian noise of slowly varying means and variance $(\mu_{S_i}(n), \sigma^2_{S_i}(n))$, and $A \gg \sigma_{S_i}(n)$. Accordingly, the detection process consists in the following steps. Firstly, the slowly varying means and variances $\mu_m$ and $\sigma_m$ are estimated over 251-frame long windows centered at frame $m$, $W(n) = [S_i(n-125) \ldots S_i(n) \ldots S_i(n+125)]$. Classical estimators, such as empirical mean and variance, cannot be used as the data is polluted by outliers corresponding to occurences of drum hits. A somewhat more accurate estimation can be obtained by considering the median of $W(n)$ as an estimate of $\mu_{S_i}(n)$; and the trimmed variance (which consists in rejecting the 10% highest and lowest values of $W(n)$ before using the classical estimator) as an estimate of $\sigma^2_{S_i}(n)$.

Thus, the normalized detection function $S_i'(n) = \frac{S_i(n) - \mu_{S_i}(n)}{\sigma_{S_i}(n)}$ can be considered as a centered gaussian process of unitary variance, with pulses $r_{S_i}(n)$ superposed. The detection can be achieved by filtering $S_i'(n)$ by the matched filter of impulse response $r_{S_i}(-n)$, yielding $S_i''(n)$, and by considering the probability that an observed sample cannot be accounted by the background noise $d_{S_i}(n) = 1 - \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-S_i''(n)^2}{2}\right)$.

This procedure produces for each region $\mathscr{R}_i$ two sequences of probabilities $d_{S_i}(n)$ and $d_{M_i}(n)$, respectively expressing the probability that the drum sticks intersect the region $\mathscr{R}_i$ at frame $n$, and the probability that the instrument corresponding to $\mathscr{R}_i$ is moving. A drum hit should be detected when both of these events occur simultaneously. Thus, a conjunction rule (product) is used to aggregate these probabilities, in order to obtain the posterior probability $p_i(n)$ that the instrument corresponding to $\mathscr{R}_i$ has been played at frame $n$. $p_i(n)$ can then be compared to a threshold to obtain the sequence of frame indices (or instants) $H_i^{video}$ at which the instrument associated to $\mathscr{R}_i$ is hit.

## 5. MAPPING REGIONS TO INSTRUMENT CLASSES

In order to achieve a complete transcription of the performance, the next step consists in labeling each region with a symbol indicating the specific instrument of the drum kit (snare drum, hi-hat, etc.) it contains. Let $\mathscr{L}_j$ be the $j$th candidate label ($j \in [1, N_L]$). The labeling task consists in finding a mapping $\varphi$ between the set of regions and the set of labels. This task proves to be very challenging as the material used for the evaluation contains a number of situations (left-handed drummer, afro or salsa rhythms played mostly on the toms rather than on the snare drums) defeating simple heuristics based on the position of the regions with respect to the drummer, or the frequency of hits. We consequently consider more robust clues to perform this mapping: the color of the drum elements, and the consistency of the events detected from the video stream with a (partial) transcription obtained by an audio drum event detection system.

### 5.1 Color criterion

The instruments in $\mathscr{L}_j$ fall into two broad categories: drums (snare drum and several sizes of toms), and cymbals (hi-hat, ride, crash or splash cymbals). Let $\mathscr{C}(\mathscr{L}_j)$ be the category to which the label $\mathscr{L}_j$ belongs.

Cymbals are made from copper-based alloys and can be identified on the image from their color. Thus, for each detected region $\mathscr{R}_i$, 16-bin hue, saturation and value histograms are computed, yielding a feature vector $X_i$ of dimension 48. A Support Vector Machine classifier using a Gaussian kernel is trained to discriminate the elements of the drum kit into these two classes. Let $\mathscr{C}(\mathscr{R}_i)$ the

category assigned to the region $\mathscr{R}_i$ by this automatic classification process. A compatibility matrix between labels and regions can thus be defined as:

$$C_{i,j}^{color} = \delta_{\mathscr{C}(\mathscr{R}_i)}^{\mathscr{C}(\mathscr{L}_j)}$$

An example for such compatibility matrix with 4 labels and 2 regions is given in the table below.

| Label $L_j$ | Region $R_i$ $\mathscr{C}(\mathscr{L}_j)$ \ $\mathscr{C}(\mathscr{R}_i)$ | $R_1$ Drum | $R_2$ Cymbal |
|---|---|---|---|
| Snare drum | Drum | 1 | 0 |
| Hi-hat | Cymbal | 0 | 1 |
| Crash cymbal | Cymbal | 0 | 1 |
| Medium Tom | Drum | 1 | 0 |

### 5.2 Consistency criterion

We assume at this stage that another transcription system, which does not rely on the video modality, is available. This assumption is realistic in applications where visual analysis is used to enhance an existing audio drum transcription system of limited accuracy. Such transcription systems (As described for example in [21, 22]) produce, for each instrument category $\mathscr{L}_j$, the sequence $H_j^{audio}$ of instants at which the instrument $\mathscr{L}_j$ is played. If a region $\mathscr{R}_i$ is associated to the instrument $\mathscr{L}_j$, we expect the sequences of detected events $H_i^{video}$ and $H_j^{audio}$ to be consistent, and to contain mostly elements common to each other. In order to measure this consistency, we propose the following criterion:

$$C_{i,j}^{cons} = \frac{|H_i^{video} \cap H_j^{audio}|}{\sqrt{|H_i^{video}|}\sqrt{|H_j^{audio}|}}$$

where $|\cdot|$ denotes set cardinality. This criterion can be seen either as the number of co-occurences normalized by the geometric mean of the number of events detected from each modality, or as a binary version of the Pearson correlation between audio and video detection functions. Because the audio and video classifiers employ different time resolutions, and because they can detect the same event with a slight delay, the elements of $H_i^{video}$ and $H_i^{audio}$ are quantized on a 100 ms uniformly spaced grid.

### 5.3 Finding an optimal mapping

The two compatibility matrices previously obtained can be combined to yield a global compatibility matrix between the labels and regions $C_{i,j} = C_{i,j}^{color} + C_{i,j}^{cons}$. The optimal mapping $\varphi^*$ between the regions and labels is the one yielding the maximum total compatibility, that is to say:

$$\varphi^* = \arg\max_{\varphi} \sum_i C_{i,\varphi(i)}$$

This problem can be reformulated as finding a maximum weight matching in a bipartite graph whose edge weights are defined by the compatibility matrix $w(e) = C_{i,j}$ if $e = \{\mathscr{R}_i, \mathscr{L}_j\}$. The Kuhn-Munkres algorithm [23] can efficiently solve this problem in $\mathbf{O}(N^3)$ where $N = \max\{N_L, N_R\}$.

## 6. FUSION WITH OTHER DETECTORS

The video transcription process described in the previous sections produces, for each instrument class $\mathscr{L}_j$ the sequence $p_j^1(n) = p_{\phi^{-1}(j)}(n)$ of probabilities that this instrument is played at frame $n$. When the scene is recorded by more than one video camera, the analysis can be individually performed for each video stream. Additionally, an audio transcription system can also be used. If $N_s$ streams are considered, a final decision can be taken by fusing their detectors' outputs $p_j^1(n) \dots p_j^{N_s}(n)$. Assuming that the information brought by each modality is accurate and complementary, a disjunctive rule can be used to achieve this fusion:

$$p_j(n) = 1 - \prod_{i=1}^{N_s} (1 - p_j^i(n))$$

Figure 2: Example of drum top extraction

Table 1: Drum top detection performances (precision and recall) for each drummer and camera angle

| **Drummer** | **1** | | **2** | | **3** | |
|---|---|---|---|---|---|---|
| **Angle** | **1** | **2** | **1** | **2** | **1** | **2** |
| % Precision | 100 | 100 | 73 | 83 | 100 | 100 |
| % Recall | 83 | 100 | 67 | 56 | 37 | 90 |

## 7. EXPERIMENTAL RESULTS

The evaluation was conducted on 51 video sequences from the public evaluation database [24]. Each sequence is captured from 2 different camera angles. The sequences were played by 3 drummers, each of them playing on a different kit, on top of a background music accompaniment – an adverse situation for audio-only drum transcription systems. The average sequence length is 60 seconds, and each sequence contains an average of 520 drum events (hits), all manually annotated. The taxonomy uses different labels for each kind of cymbal (splash, crash, ride) and each tom size (low, medium, floor toms) – again information that is difficult to extract from the audio signal only.

In order to evaluate the ellipse detection algorithm, one sequence was selected for each drummer and angle. The frames from these sequences were averaged in advance in order to reduce occlusions caused by the drummer. Table 1 presents the precision and recall values obtained. Figure 2 displays an example of the output of the procedure, with the detected drum tops highlighted.

Then, the entire drum transcription system described in figure 1 is evaluated. Note that the ellipse detection is, in this case, performed on the average frame. As an intermediary result, table 2 presents the accuracy of the automatic region labeling process described in section 5. Transcription results, obtained from several sources of information (Audio, Video camera 1, Video camera 2) are presented in table 3. The F-measure, which expresses a trade-off between precision and recall, is used as a performance measure. For comparison, transcription results using a semi-automatic system (in which the drum top detection and region labeling processes are performed by human operators) are also given.

It can be seen that visual analysis is better than the audio transcription system at identifying toms and cymbal hits. Acceptable performances can be achieved for the detection of snare drums and hi-hat hits, though the detection is not as accurate as on the audio

Table 2: Instrument identification accuracy for each drummer and camera angle

| **Drummer** | **1** | | **2** | | **3** | |
|---|---|---|---|---|---|---|
| **Angle** | **1** | **2** | **1** | **2** | **1** | **2** |
| % Accuracy | 64 | 76 | 60 | 73 | 72 | 64 |

Table 3: Transcription accuracy (% F-measure) for each instrument of the drum kit, using various combinations of the audio and video streams

| | Unimodal | | | | Multimodal | | |
|---|---|---|---|---|---|---|---|
| Audio | • | | | | • | • | • |
| Video angle 1 | | • | | • | • | | • |
| Video angle 2 | | | • | • | | • | • |
| **Manual segmentation and labeling** | | | | | | | |
| Bass drum | **69** | 40 | 0 | 40 | 66 | **69** | 66 |
| Snare drum | 63 | 52 | 45 | 53 | **72** | 69 | 71 |
| Toms | 7 | 11 | 14 | 14 | 11 | 18 | **20** |
| Hi-hat | 78 | 63 | 67 | 70 | 75 | **79** | 77 |
| Cymbals | 17 | 27 | **28** | **28** | 23 | 26 | **28** |
| **Automatic segmentation and labeling** | | | | | | | |
| Bass drum | **69** | 0 | 0 | 0 | **69** | **69** | **69** |
| Snare drum | 63 | 49 | 36 | 51 | **69** | 66 | **69** |
| Toms | 7 | 12 | 21 | **22** | 16 | 18 | 18 |
| Hi-hat | **78** | 48 | 63 | 66 | 77 | 77 | 77 |
| Cymbals | 17 | 29 | 29 | **30** | 27 | 27 | 29 |

signal. This complementarity is well-exploited by the disjunctive fusion rule, whose performances per instrument, when manual segmentation is used, are similar or better than the best of the unimodal approaches. Interestingly, the automatic approach can yield in some cases better results than the manual annotation – some of the misaligned or spurious ellipses obtained by the segmentation process can allow a more precise detection that an ellipse precisely fitting the edges of the drum top. This suggests that the most relevant regions of interest to detect drum hits might not be the drum top themselves, the regions could rather be extended to include the stick and drummer forearms.

## 8. CONCLUSION AND FUTURE WORK

This paper presented a video drum transcription system, with an emphasis on the image segmentation front-end. This system, which requires no calibration and no prior training, has been tested on a large and diverse database, and showed improvement over audio-only approaches for the detection of tom and cymbal hits. A fusion of the audio and video transcriptions is possible with a disjunction rule, as both modalities provide complementary information. Future work will investigate the use of multimodal approaches at the feature level, rather than at the decision level. Especially, the extraction of individual audio/visual components by Non-negative Matrix Factorization could provide both a segmentation of the image into non-geometrically constrained regions of interest, and detection functions for each component.

## 9. ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of the European Commission under the FP6-027026-K-SPACE contract.

## REFERENCES

[1] A. Klapuri and M. Davy, editors. *Signal Processing methods for the automatic transcription of music*. Springer, 2006.

[2] D. Murphy. Tracking a conductor's baton. In *Proc. 12th Danish Conf. on Pat. Recognition and Image Ana.*, 2003.

[3] M. M. Wanderley and P. Depalle. Gesturally-controlled digital audio effects. In *Proc. 5th Int. Conf. on Digital Audio Effects (DAFX'02)*, December 2001.

[4] S. Dahl. The playing of an accent - preliminary observations from temporal and kinematic ana. of percussionists. In *Journal of New Music Research*, volume 29(3), pages 225–234, 2000.

[5] J. Saitoh, A. Kodata, and H. Tominaga. Integrated data processing between image and audio - musical instrument (piano) playing information processing. In *Proc. 6th Int. Conf. on Image Proc. and its Applications*, 1997.

[6] O. Gillet and G. Richard. Automatic transcription of drum sequences using audiovisual features. In *Proc. 2005 IEEE Conf. on Acoustics, Speech and Signal Proc. (ICASSP'04)*, 2005.

[7] J Canny. A computational approach to edge detection. *IEEE Trans. Pat. Ana. Mach. Intel.*, 8(6):679–698, 1986.

[8] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. *ICCV*, 00:839, 1998.

[9] G. Taubin. Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applications to edge and range image segmentation. *IEEE Trans. Pat. Ana. and Mach. Intel.*, 13(11):1115–1138, Nov 1991.

[10] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[11] Tsuyoshi Kawaguchi and Ryo ich Nagata. Ellipse detection using a genetic algorithm. *ICPR*, 01:141, 1998.

[12] A. Fitzgibbon, M. Pilu, and R.B. Fisher. Direct least square fitting of ellipses. *IEEE Trans. Pat. Ana. and Mach. Intel.*, 21(5):476–480, May 1999.

[13] C. Bond. A new algorithm for scan conversion of a general ellipse. January 2002.

[14] David Eberly. Distance from a point to an ellipse in 2d. http://www.geometrictools.com/, October 2002.

[15] Pedro F. Daniel and Daniel P. Huttenlocher. Distance transforms of sampled functions. Cornell Computing and Information Science Technical Report, September 2004.

[16] Joshua Podolak, Philip Shilane, Aleksey Golovinskiy, Szymon Rusinkiewicz, and Thomas Funkhouser. A planar-reflective symmetry transform for 3d shapes. *ACM Trans. Graph.*, 25(3):549–559, 2006.

[17] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Symmetry descriptors and 3d shape matching. In *SGP '04: Proc. 2004 Eurographics/ACM SIGGRAPH symposium on Geometry Proc.* ACM Press, 2004.

[18] J. Ross Quinlan. *C4.5: Programs for Mach. Learning*. Morgan Kaufmann Publishers, 1993.

[19] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. 1999 IEEE Computer Society Conf. on Computer Vision and Pat. Recognition (CVPR'99)*, 1999.

[20] Maxim Dvornikov. Formulae of numerical differentiation. math.NA/0306092, 2003.

[21] O. Gillet and G. Richard. Drum track transcription of polyphonic music using noise subspace projection. In *Proc. 6th Int. Conf. on Music Info. Retrieval (ISMIR'05)*, September 2005.

[22] K. Tanghe, S. Degroeve, and B. De Baets. An algorithm for detecting and labeling drum events in polyphonic music. In *Proc. 2005 MIREX evaluation campaign*, 2005.

[23] H. W. Kuhn. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[24] O. Gillet and G. Richard. ENST-drums: an extensive audio-visual database for drum signals processing. In *Proc. 7th Int. Conf. on Music Info. Retrieval (ISMIR'06)*, 2006.