Gyro-based camera-motion detection in user-generated videos

Sophia Bano[*†], Andrea Cavallaro[*], Xavier Parra[†]
[*]Centre for Intelligent Sensing, Queen Mary University of London, United Kingdom.
[†]Technical Research Centre for Dependency Care & Autonomous Living,
Universitat Politècnica de Catalunya, Spain.
{s.bano, a.cavallaro}@qmul.ac.uk; xavier.parra@upc.edu

ABSTRACT

We propose a gyro-based camera-motion detection method for videos captured with smartphones. First, the delay between the acquisition of video and gyroscope data is estimated using similarities induced by camera motion in the two sensor modalities. Pan, tilt and shake are then detected using the dominant motions and high frequencies in the gyroscope data. Morphological operations are applied to remove outliers and to identify segments with continuous camera-motion. We compare the proposed method with existing methods that use visual or inertial sensor data.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis, Motion*

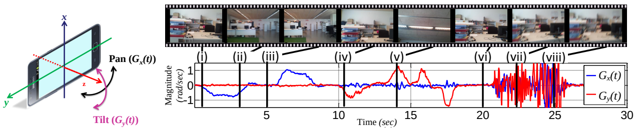
Keywords

Gyroscope sensor, Visual sensor, User-generated videos, Camera motion, Detection, Synchronization

1. INTRODUCTION

Camera-Motion Detection (CMD) helps video summarization, composition, shot detection, segmentation and content analysis [1, 3, 4, 12]. When available, inertial sensors complement visual data and can support video-quality analysis with modest memory and battery consumption. We are interested in detecting camera motion that a user introduces while recording a video with a hand-held camera.

Most CMD methods analyze visual content, for example using template matching [1, 8], optical flow [2, 9, 10] or Luminance Projection Correlation (LPC) [4, 11]. Template matching can be applied to estimate the horizontal, vertical and zoom motion combined with a support vector machine [1, 8]. Optical flow-based methods classify frames into stationary or moving frames based on the magnitude of motion vectors. Moving frames are then classified into pan, tilt or zoom frames based on the dominant orientation or template matching [2, 9, 10]. LPC [11] correlates the horizontal and vertical projections of consecutive frames to compute pan and tilt [3, 4, 12, 13]. Existing methods are generally computationally expensive and can be corrupted by moving objects and brightness changes. Inertial sensors can help overcome these problems.



(a) (b)

Figure 1: (a) Definition of a device reference coordinates, pan $G_x(t)$ and tilt $G_y(t)$ signals. (b) A sample multi-modal video containing pan, tilt and shake: (i) and (ii): pan; (iv) and (v): tilt; (vii) and (viii): shake.

Cricri *et al.* [5] use inertial sensor data for event understanding in User-Generated Videos (UGVs). Low-Pass Filtered (LPF) compass data (sampled at $10Hz$) are used to detect pan by calculating the angular speed of the camera; the tilt angle is computed from unfiltered accelerometer data (sampled at $40Hz$); and High-Pass Filtered (HPF) accelerometer data are analyzed to detect shake. However, compass is sensitive to drift and errors induced by nearby magnetic objects, and unfiltered accelerometer data can affected by noise that reduces the detection accuracy.

Unlike accelerometer and compass, a gyroscope is neither affected by gravity nor magnetic field. A gyroscope measures the angular velocities around the device's x, y and z axes (Fig. 1(a)). These velocities correspond to camera pan, tilt and roll, respectively. A high correlation on the motion of a device exists between visual and gyroscope data captured from the same device.

In this paper, we propose a gyro-based method for CMD. To the best of our knowledge, this is the first method that uses the output of a gyroscope to detect pan, tilt and shake in UGVs. We exploit the tri-axial gyroscope data captured along with the video to identify visual frames with motion. Because visual data logged with an inertial sensor have unknown delay due to the time taken for the camera initializa-

S. Bano was supported by the EACEA Agency of the European Commission under EMJD ICE FPA n° 2010-0012. A. Cavallaro acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC), under grant EP/K007491/1.

MM'15, October 26–30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806336.

tion, we synchronize video and gyroscope data using similarities induced by the camera motion itself. Dominant motions of the LPF gyroscope data in polar coordinates are utilized for pan and tilt detection, whereas dominant high-frequency motions are considered for shake detection. We further apply morphology to remove outliers and to identify segments with the same camera-motion type.

2. PROPOSED METHOD

Let $C(t) = \{V(t), D(t)\}$ be the multi-modal data of a UGV captured with a hand-held device. $V(t)$ is the visual data sampled at V_r frames per second (*fps*) and $D(t)$ represents the inertial data stream sampled at f_r *Hz*. $D(t)$ contains tri-axial gyroscope data $G(t) = (G_x(t), G_y(t), G_z(t))$ logged *simultaneously* with $V(t)$.

We aim to estimate the delay $\hat{t}$ for gyro-visual synchronization, and then to detect pan $P_d(t)$, tilt $T_d(t)$ and shake $S_d(t)$ using gyroscope data only. We assume that the gyroscope is auto-calibrated such that $G_x(t)$, $G_y(t)$ and $G_z(t)$ are zero at steady state and the gyroscope data are free from offset error.

2.1 Gyro-visual synchronization

The magnitude of $G(t)$ for a smartphone video captured without a tripod is affected by involuntary human body movements. Even in the absence of intentional camera motion, this low-magnitude motion due to involuntary body movement is sufficient for gyro-visual data synchronization. The frequency of involuntary human body movements when holding a device lies within $f_i = 20Hz$ [7]. Inertial sensors are logged and analyzed at $f_r = 50Hz > 2f_i$, thus satisfying the Nyquist theorem [6].

We correlate gyroscope and visual data to determine the recording delay between the two modalities. Firstly, LPC [11] (detailed in [13]) is used to compute the horizontal $L_x(t)$ and vertical $L_y(t)$ displacements from $V(t)$. $L_x(t)$ and $L_y(t)$ are the video-based pan and tilt, and correspond to the gyroscope $G_x(t)$ and $G_y(t)$ signals, respectively.

Let $I(t,i,j)$ be the grayscale value of pixel (i,j) at time t, and $I_y(t,i)$ and $I_x(t,j)$ its horizontal and vertical projections, respectively. These projections are computed as

$$I_y(t,i) = \frac{1}{h}\sum_{j=1}^{h} I(t,i,j)\ , \quad I_x(t,j) = \frac{1}{w}\sum_{i=1}^{w} I(t,i,j), \tag{1}$$

where h is the height and w is the width of the frame.

$L_x(t)$ and $L_y(t)$ (Fig. 2(a)(b)) are then computed as

$$L_x(t) = arg\ \underset{\delta p}{min} \sum_{i=1+\delta p_1}^{w-\delta p_2} D_{I_y}(t,i,\delta p), \tag{2}$$

$$L_y(t) = arg\ \underset{\delta p}{min} \sum_{j=1+\delta p_1}^{h-\delta p_2} D_{I_x}(t,j,\delta p), \tag{3}$$

where $\delta p \in [-20, 20]$. When $\delta p < 0$ then $\delta p_1 = 0$, $\delta p_2 = \delta p$, whereas when $\delta p \geq 0$ then $\delta p_1 = \delta p$, $\delta p_2 = 0$. $D_{I_x}(t,j,\delta p)$ and $D_{I_y}(t,i,\delta p)$ are computed as follows:

$$D_{I_x}(t,j,\delta p) = (I_x(t,j) - I_x(t+1, j-\delta p))^2,$$
$$D_{I_y}(t,i,\delta p) = (I_y(t,i) - I_y(t+1, i-\delta p))^2,$$

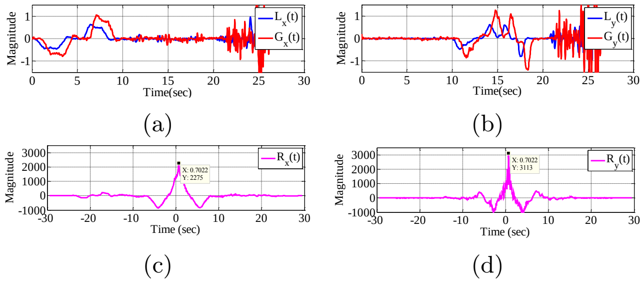


Figure 2: Gyro-visual synchronization. (a) LPC pan $L_x(t)$ and gyroscope $G_x(t)$, (b) LPC tilt $L_y(t)$ and gyroscope $G_y(t)$. Correlation (c) $R_x(t)$ and (d) $R_y(t)$.

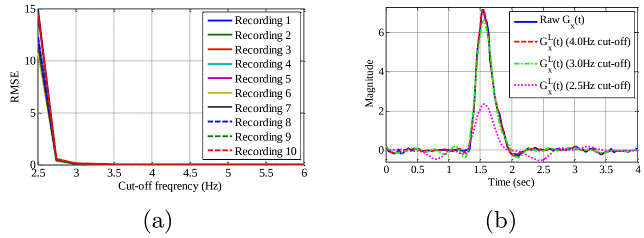


Figure 3: Analysis of f_c for the LPF. (a) The RMSE between the $G_x(t)$ and $G_x^L(t)$ for 10 recordings containing fast pan for varying values of f_c. (b) The raw $G_x(t)$ (for the fast pan) for one of the recording and its LPF signals with varying f_c.

where $D_{I_x}(.)$ and $D_{I_y}(.)$ are the projection distances between two consecutive frames in the horizontal and vertical direction, respectively.

We downsample $G(t)$ to have the same sampling rate as $V(t)$ and compute the pan cross-correlation

$$R_x(t) = \sum_{k=-\infty}^{\infty} G_x(k) L_x(k+t). \tag{4}$$

The correlation peak is $\hat{e}_x = \max_t R_x(t)$ and the estimated delay is $\hat{t}_x = \text{argmax}_t R_x(t)$. Likewise, $R_y(t)$, $\hat{t}_y$ and $\hat{e}_y$ are computed from the tilt signals. The overall estimated delay $\hat{t}$ is selected as

$$\hat{t} = \begin{cases} \hat{t}_x & if\ \hat{e}_x \geq \hat{e}_y, \\ \hat{t}_y & if\ \hat{e}_y > \hat{e}_x. \end{cases} \tag{5}$$

Figure 2(c)(d) shows an example of cross correlations $R_x(t)$ and $R_y(t)$, and their corresponding peaks.

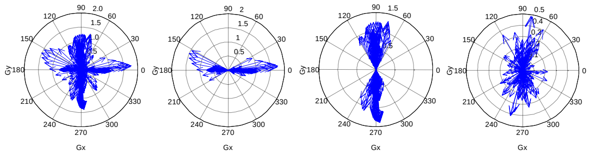
2.2 Gyro-based camera motion detection

Panning produces higher magnitudes of $G_x(t)$ whereas tilting produces higher magnitudes of $G_y(t)$. We low-pass filter the signals to obtain $G_x^L(t)$ and $G_x^L(t)$ using a cut-off frequency $f_c = 4Hz$ (Fig. 3).

We then jointly analyze $G_x^L(t)$ and $G_y^L(t)$ in polar coordinates with radial value $G_r^L(t) = \sqrt{G_x^L(t)^2 + G_y^L(t)^2}$ and angular value $G_\theta^L(t) = \arctan\left(\frac{G_y^L(t)}{G_x^L(t)}\right)$ (Fig. 4).

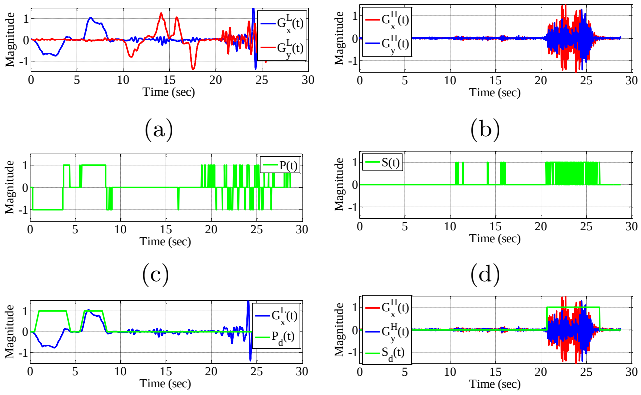
Pan, $P(t)$, is detected as

$$P(t) = \begin{cases} +1 & if\ 0-\alpha \leq G_\theta^L(t) \leq 0+\alpha, \\ -1 & if\ 180-\alpha \leq G_\theta^L(t) \leq 180+\alpha, \\ \ \ 0 & otherwise, \end{cases} \tag{6}$$



(a) (b) (c) (d)

Figure 4: Analysis of $G_x(t)$ and $G_y(t)$ in polar coordinate. (a) $G_r^L(t)$ and $G_\theta^L(t)$, detected (b) pan and (c) tilt vectors, (d) $G_r^H(t)$ and $G_\theta^H(t)$ for shake detection are shown.



(a) (b) (c) (d) (e) (f)

Figure 5: Camera motion detection. For pan, (a) $G_x^L(t)$ and $G_y^L(t)$, (c) $P(t)$, (e) $P_d(t)$, and for shake, (b) $G_x^H(t)$ and $G_y^H(t)$, (d) $S(t)$, (f) $S_d(t)$ are shown.

where α (*degrees*) is a tolerance angle, and +1 and −1 denote pan left and right, respectively (Fig. 5c). Similarly, tilt $T(t)$ is detected as

$$T(t) = \begin{cases} +1 & \textit{if } 90-\alpha \leq G_\theta^L(t) \leq 90+\alpha, \\ -1 & \textit{if } 270-\alpha \leq G_\theta^L(t) \leq 270+\alpha, \\ \ \ 0 & \textit{otherwise}, \end{cases} \qquad (7)$$

where +1 and −1 denote tilt down and up, respectively. The value of α should facilitate the detection of horizontal and vertical motions. Because freehand pan is not strictly along the x-axis in real-world scenarios, a higher magnitude of $G_x(t)$ is also associated to a non-zero value of $G_y(t)$ while panning with a smartphone. We therefore empirically set $\alpha = 30^o$ (Fig. 6(a)).

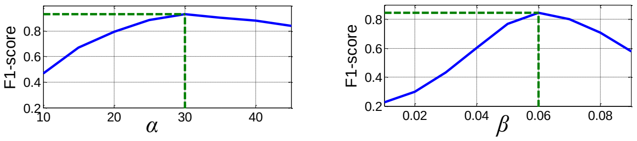
For shake detection, we obtain $G_r^H(t)$ and $G_\theta^H(t)$ after high-pass filtering $G_x^H(t)$ and $G_y^H(t)$ (Fig. 4(d)). $G_r^H(t)$ is used to classify shake camera motion as

$$S(t) = \begin{cases} 1 & \textit{if } G_r^H(t) > \beta, \\ 0 & \textit{otherwise}. \end{cases} \qquad (8)$$

$G_r^H(t)$ ranges from 0 to 0.5. We selected $\beta = 0.06$ empirically (Fig. 6(b)) to remove the influence of involuntary body movement in freehand recordings.

$|P(t)|$, $|T(t)|$ and $S(t)$ give binary signals for samples with detected pan, tilt and shake motions (Fig. 5(c-d)).

As we have not yet considered time continuity, false camera motion can be detected in few samples. To remove these outliers and to identify time-continuous camera-motion segments, we apply morphological operations to the binary signals [6]. In particular, we apply *opening* to $|P(t)|$ and $|T(t)|$ to remove false detection by erosion, followed by dilation to detect continuous segments, and to obtain the final motion-detection results $P_d(t)$ and $T_d(t)$ (Fig. 5(e)). To detect continuous segments of shake, we apply *closing* that performs dilation to connect discontinuous segments followed by erosion to maintain the original length of the shake detected segments. Segments that are shorter than 0.25 seconds are considered as outliers, and are removed to obtain the final shake detection signal $S_d(t)$ (Fig. 5(f)).



(a) (b)

Figure 6: F$_1$-score of the proposed CMDG method with respect to the varying values of (a) α and (b) β for the captured multi-modal dataset.

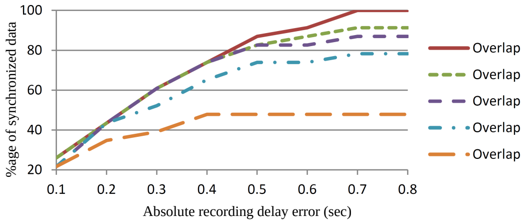


Figure 7: Gyroscope-visual synchronization result for 24 multi-modal UGVs. %age of synchronized data *w.r.t* the absolute delay error for varying Overlap$N$.

3. EXPERIMENTAL EVALUATION

To evaluate the proposed method, CMDG, we recorded multi-modal data in several scenarios such as concerts, parades, festivals and firework display, using Cellbots Data Logger[1] and compared with an existing visual [13, 4] (referred as VISUAL) and inertial sensor-based [5] (referred as ISENSOR) method.

We captured 24 UGVs for a total duration of 70 *mins* in High Brightness (HB) and Low Brightness (LB) scenarios (e.g. day and night-time). This collected dataset is available for downloading[2] and it contains single camera recordings at distinct timings and locations, changing lights and varying camera motions. The Ground-Truth (GT) delay is obtained by observing a pan/tilt/shake motion simultaneously in the gyro-visual data. Each captured video was manually annotated to get labels for camera motions for each second.

Figure 7 shows the results for gyro-visual synchronization for the captured multi-modal dataset. Acceptable delays are obtained for all UGVs with an absolute error of 0.7*sec* between GT and estimate. This error is mainly due to the imprecise GT labels as it was difficult to manually observe a coherent motion both in the video and gyroscope data. By

[1] `https://cellbots.googlecode.com/files/CellbotsDataLogger_v1.1.0_full.apk` (Last accessed: 30/07/2015). Note that the video frame rate and frame size varied depending on the brightness of the recorded scene due to the programmed settings of the App.

[2] `http://www.eecs.qmul.ac.uk/~andrea/cmdg`

Table 1: Results for CMDG and its comparison with a VISUAL [13, 4] and ISENSOR [5] method. Key: HB: high brightness recordings; LB: low brightness recordings; TP: true positive; FP: false positive; P: precision; R: recall; F_1: F_1 score.

		Pan						**Tilt**						**Shake**					
Method	**Type**	**GT**	**TP**	**FP**	**P**	**R**	**F_1**	**GT**	**TP**	**FP**	**P**	**R**	**F_1**	**GT**	**TP**	**FP**	**P**	**R**	**F_1**
CMDG	HB	294	272	11	0.96	0.93	**0.94**	36	29	6	0.83	0.81	**0.82**	389	365	129	0.74	0.94	**0.83**
VISUAL [13, 4]			217	64	0.77	0.74	0.75		19	42	0.31	0.53	0.39		260	118	0.69	0.67	0.68
ISENSOR [5]			175	52	0.77	0.60	0.67		14	46	0.23	0.39	0.29		188	93	0.67	0.48	0.56
CMDG	LB	123	117	12	0.91	0.95	**0.93**	49	41	7	0.85	0.84	**0.85**	272	235	37	0.86	0.86	**0.86**
VISUAL [13, 4]			31	44	0.41	0.25	0.31		10	48	0.17	0.20	0.19		200	606	0.25	0.74	0.37
ISENSOR [5]			44	40	0.52	0.36	0.43		23	24	0.49	0.47	0.48		213	129	0.62	0.78	0.69

jointly visualizing the synchronized data, we cross-validated the correctness of the obtained results. Acceptable delay is achieved even in the presence of slight camera motion. Although the illumination is extremely low in some LB recordings (e.g. fireworks) that resulted in low magnitude of $L_x(t)$ and $L_y(t)$ and inaccurate CMD, correlation existed between the gyro-visual data. A slight clue of brightness (e.g. an exploding firework) is sufficient to establish the correlation.

To investigate the robustness of gyro-visual synchronization, we changed the overlap duration between the gyroscope and visual data for the whole dataset. Overlap$N$ denotes that complete visual data and only N% of the duration of the gyroscope data are used. For Overlap80, Overlap60, Overlap40 and Overlap20, the percentage of synchronized recordings are 91%, 87%, 78% and 48%, respectively (Fig. 7). Note that the visual quality and frame rate are low in some of the night-time recordings: this affects $L_x(t)$ and $L_y(t)$, and decreases the performance when the overlap decreases.

For the evaluation of the proposed CMDG, we analyze its performance with respect to the GT and compared with alternative approaches (Table 1). To have a fair comparison, the parameters within the VISUAL and ISENSOR are adjusted to give the best possible results. To select α and β for pan, tilt and shake detection, we analyzed the effect of varying these parameters on the detection results (Fig. 6), for the captured multi-modal dataset. At $\alpha = 30^o$, the best F_1-score of 0.94 for pan and tilt, and at $\beta = 0.06$, the best F_1-score of 0.85 for shake were achieved and selected for the experimentation. In order to investigate their performance, we divided the UGVs into HB and LB recordings having total durations of 30 *mins* and 40 *mins*, respectively. In our dataset, most events of interest existed in the latitudinal plane (e.g. singer, crowd, parade), with the exception of few that existed in the longitudinal plane (e.g. fireworks, flying balloons), resulting in fewer tilt samples (Table 1). CMDG outperformed the existing methods giving the F_1-score of 94%, 82% and 83% for $P_d(t)$, $T_d(t)$ and $S_d(t)$, respectively, for the HB recordings, and 93%, 85% and 86% for the LB recordings. VISUAL and ISENSOR are the second best for the HB and LB recordings, respectively.

VISUAL is affected by the motion of objects and light conditions, thus reducing its performance in LB recordings compared to CMDG and ISENSOR, which are independent of these factors. ISENSOR uses compass and accelerometer, and is affected by magnetic noise, low sampling rate and unfiltered processing, resulting in false detections. CMDG is a more desirable solution for CMD because of the use of a more accurate sensor (gyroscope), and inclusion of the post-processing stage that suppresses outliers. Pan signals from CMDG and VISUAL are comparable in HB recordings. However, ISENSOR is less accurate due to the lower sampling rate ($10Hz$) of the compass [5]. Increasing the sampling rate to $50Hz$ increases the effect of noise that affects the derivative of the compass signal.

4. CONCLUSION

We proposed a method for camera motion detection using gyroscope data for user-generated videos. The method aligns multi-modal data and uses the tri-axial gyroscope data captured simultaneously with the video to detect pan, tilt and shake. Our proposed method outperformed existing approaches with a collective F_1-score of 89% for pan, tilt and shake.

As future work, we are interested in developing a real-time application of the proposed method and in analyzing translational motion generated when users change their position while recording.

5. REFERENCES

[1] G. Abdollahian, C. Taskiran, Z. Pizlo, and E. Delp. Camera motion-based analysis of user generated video. *IEEE Transactions on Multimedia*, 12:28–41, 2010.

[2] J. Almeida, R. Minetto, T. A. Almeida, R. S. Torres, and N. J. Leite. Robust estimation of camera motion using optical flow models. In *Springer Advances in Visual Computing*, pages 435–446. 2009.

[3] S. Bano and A. Cavallaro. Vicomp: composition of user-generated videos. *Multimedia tools and applications*, in press, 2015.

[4] M. Campanella, H. Weda, and M. Barbieri. Edit while watching: home video editing made easy. In *SPIE Multimedia Content Access*, 2007.

[5] F. Cricri, K. Dabov, I. Curcio, S. Mate, and M. Gabbouj. Multimodal extraction of events and of information about the recording activity in user generated videos. *Multimedia tools and applications*, 70:119–158, 2012.

[6] R. Gonzalez and R. Woods. *Digital Image Processing, Chap. 4 and 9.* Pearson Education, 2008.

[7] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovell, and B. G. Celler. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE Trans. Inf. Technology in Biomedicine*, 10:156–167, 2006.

[8] D. Lan, Y. Ma, and H. Zhang. A systemic framework of camera motion analysis for home video. In *IEEE International Conference on Image Processing*, 2003.

[9] A. Mahabalagiri, K. Ozcan, and S. Velipasalar. Camera motion detection for mobile smart cameras using segmented edge-based optical flow. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2014.

[10] V. Makkapati. Robust camera pan and zoom change detection using optical flow. In *National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2008.

[11] A. Nagasaka and T. Miyatake. Real-time video mosaics using luminance-projection correlation. *IEICE Transactions on Information and Systems*, 1999.

[12] M. Saini, R. Gadde, S. Yan, and W. T. Ooi. Movimash: online mobile video mashup. In *ACM Multimedia*, 2012.

[13] K. Uehara, M. Amano, Y. Ariki, and M. Kumano. Video shooting navigation system by real-time useful shot discrimination based on video grammar. In *IEEE International Conference on Multimedia and Expo*, 2004.