Real-Time Automatic Selection of the Best Shot on Object in 4K Video Stream Based on Tracking Methods in Virtual Cropped Views

Dávid Cymbalák*, Ondrej Kainz, František Jakab
The Department of Computers and Informatics, FEEI, Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic.

* Corresponding author. Tel.:+421 55 602 7017; email: david.cymbalak@cnl.sk
Manuscript submitted December31, 2014; accepted August 3, 2015.
doi: 10.17706/ijcee.2015.7.4.275-282

Abstract: In this paper, a novel approach for real-time automatic object tracking from multiple cropped views using 4K video capturing device is presented. System itself is based on TLD object tracking algorithm. As a first ultra-high definition video and selected object tracking method is analyzed. Subsections of 4K video output from capturing device are utilized to create new, so called "virtual copped views". The developed model was rudimentary implemented and tested in the laboratory conditions and it is expected to be used in the real conditions.

Key words: Object tracking, 4K video, live streaming, best shot, virtual cameras, multi-cameras systems.

1. Introduction

Object tracking is widely covered area and focus of many studies. Survey carried out to estimate the performance of the latest algorithms may be found in [1]. Say we are interested in tracking of a lecturer or some presenter on a stage, first step is to provide some form of object tracking. In this way one user with one capturing device will be able to provide single tracking of an object. However assume presence of multiple devices, i.e. multiple users and multiple capturing devices are present and say they all are trying to cover the scene. Object of interest is usually not static but has rather dynamic characteristics, due to this capturing devices have to be physically moved (case of one device), or multiple capturing devices have to be available.

2. Ultra-High Definition Video and Object Tracking Methods

2.1. 4K Video

When referring to 4K resolution we mean any video of size up to 4000 pixels per line. 4K is rather new technology, according to [2] was 4K support announced by YouTube in July of 2010. Though 4K is up to 4000 pixels per line [3] notes that most applications refer to 4K as a frame size with dimensions of 3840x2160 which is four times full HD resolution, i.e. 1080p. Authors further note that streaming of uncompressed 4K video is difficult due to high data bit-rate. However once some form of compression is used the communication is not that critical, this was proved by [4] who developed 4K bi-directional communication system using JPEG 2000. For brief introduction to 4K technology see [5].

2.2. Object tracking in video

From the variety of object tracking approaches we primarily focus on Tracking-Learning-Detection or for short TLD algorithm which was introduced by ZdenekKalal as a part of his PhD thesis [6]. Idea of TLD was to design an algorithm for long-term tracking of the unknown objects (selection carried out by bounding box), however detection of the objects with the known pattern is available as well (image with an known object of interest). Four essential components of TLD are:

- Tracker — used for object motion estimation while having adaptive properties towards appearance and illumination changes.
- Learning — used for preserving model of an object, this component is analyzing the outputs of the one previously described component and of the detector; errors introduced by detector as estimated.
- Detector — used for detection of appearances represented by object.
- Integrator — used for incorporation of outputs from tracker and detector; its output is hypothesis about state of an object.

In order to distinguish object from the background author does not utilize any feature subtraction technique but idea of positive and negative patches that are ordered according the time. To classify which patch is which is used Nearestneighbor classification.

TLD algorithm was also used for face detection based on a priori information described in [7]. Another research focused on face detection utilizing TLD was carried out by [8]. However in this case authors developed its modified version BP-TLD, BP here stand for Back Projection which is used for calculation of histogram model of the specific feature. Authors of research [9] extended TLD of external detector and thus enabled multi-object tracking at the same time. In [10] TLD was empirically demonstrated to be used in multi-camera environment, concluding that its operation is feasible, providing lightning conditions and orientation of cameras is kept analogous.

In our research is TLD algorithm chosen to be used as a tool for the object tracking.

3. Design of Model for Selection the Best Shot on Object

3.1. Model for Selection of Best Shot for Separated Multi-camera Systems

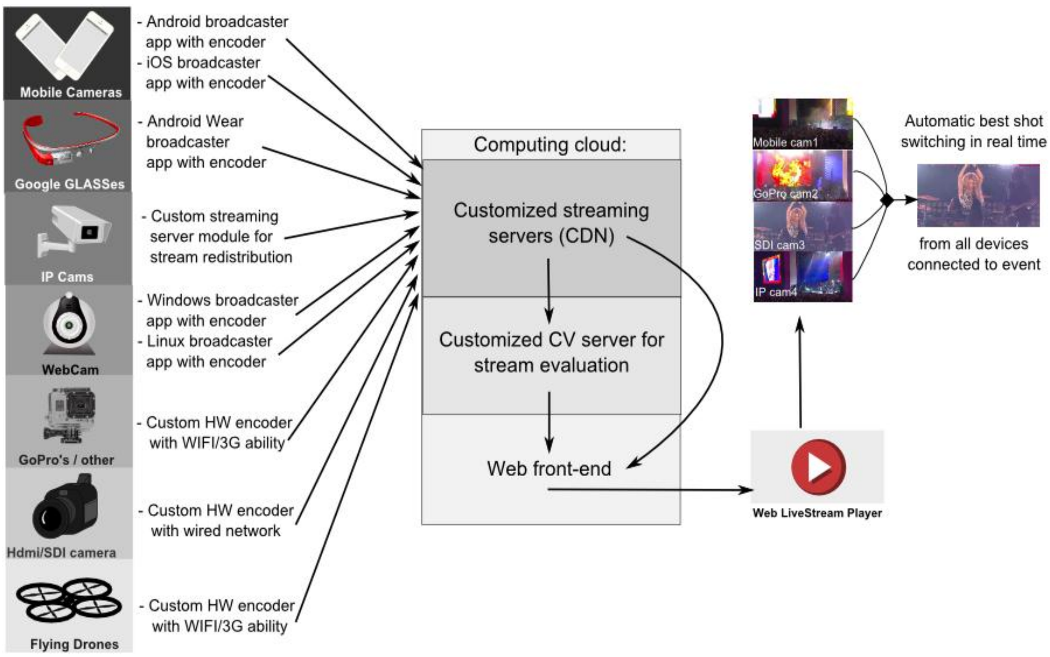

Fig. 1. Function of auto-switching solution from multiple cameras sources of different type.

Based on our previous research [11] was created model and prototypes of solution which enhances video experience from specific live streams to distant audience with autonomous director that improves surveillance and tracking abilities through the various multi-camera systems. Solution (Fig. 1) evaluates the best view on the object in real-time and delivers only the best output.

Just stated outputs of the research could also be used in cases when tracking the position of the speaker and auto-switching the best angle from multiple cameras is highly preferred. Such situations arise, for example, during conferences and university lectures/seminars. Multiple cameras, without manual operation, would track the speaker and automatically change to different viewing angles based on the movement. Our solution is designed to provide automated source switching based on actual position of the speaker without changing the physical placements of the cameras or without their rotation itself. The new approach of our auto switching best shot solution does not presume utilization of multiple physical cameras but only the one ultra-high definition camera through creation of a multiple virtual cropped views from the original image.

3.2. Creating a Virtual Cameras Views from One 4K Video Source

Frequently scene's conditions allow replacing the multiple physical cameras capturing such scene by one 4K camera, i.e. the whole scene will be captured from distance. For instance, live stream from one 4K camera capturing the conference hall from back could deliver 720p output stream composed of multiple cropped views, which brings an impression of using multiple cameras to distance viewers. Say one cropped view from whole 4K image is focused on detail of speaker, second on presentation screen, another on discussion table or the audience. This method is usable for capturing various types of scenes with only one 4K camera, while being also adaptable in surveillance or production monitoring. Previous tracking of object of interest in video enabled us to define optimal virtual camera views in 4K based on object moving trajectory (Fig. 2) or establish the most common object of appearance (Fig. 3). These virtual views could be of different sizes (360p, 480p, 720p, 1080p) and they can overlap or contain each other.

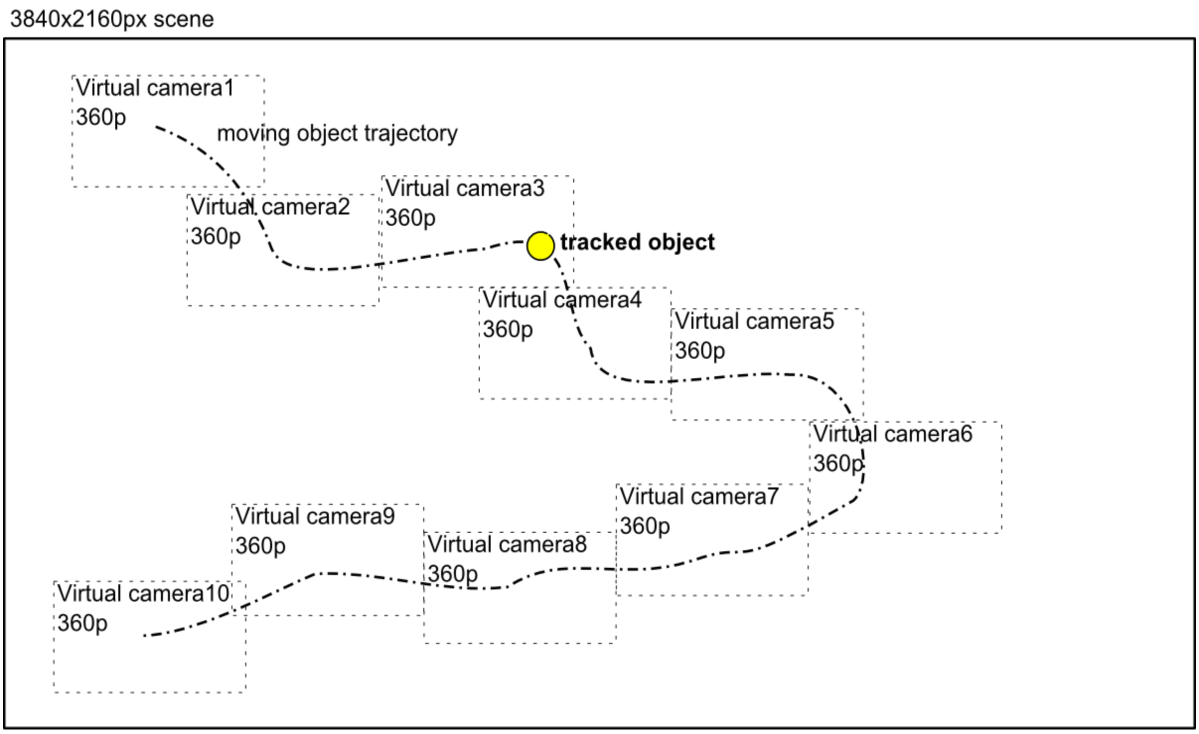

Fig. 2. Multiple virtual cameras view of size of 360p in one 4K video.

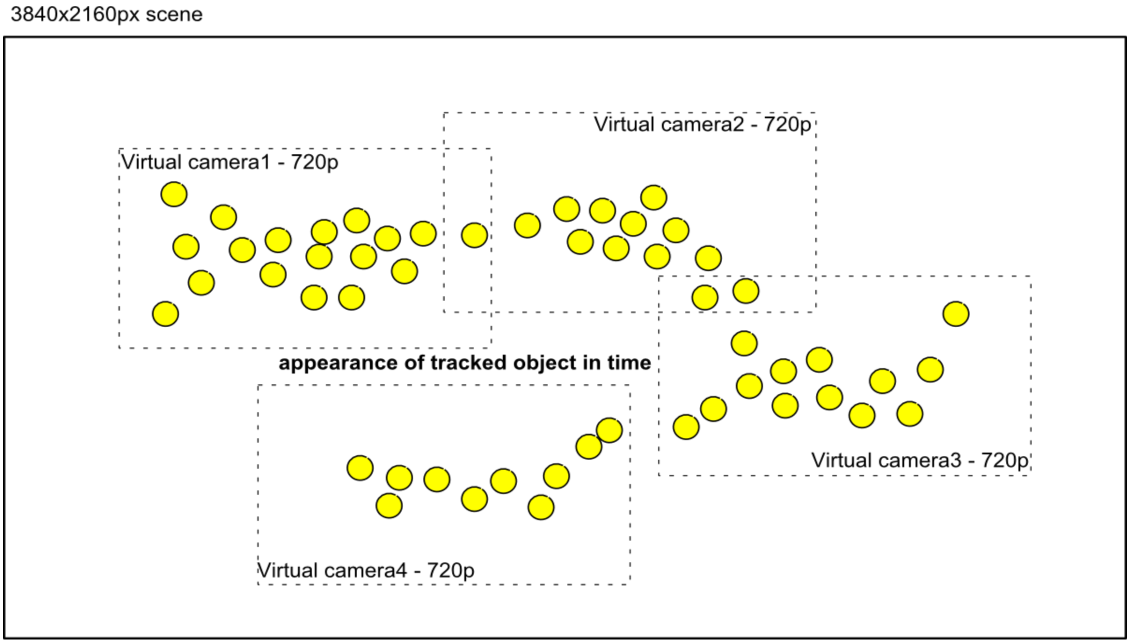

Fig. 3. Appearance of tracked object and creation of 720p virtual views.

3.3. Model for Selecting the Best Shot for Virtual Cameras in 4K Video

Once the multiple views based on previous object appearances are defined or automatically formatted there is need to use an evaluation metric for each virtual view. This metric is calculated related to parameters of tracked object from sight of each virtual view. Metric is composed of positional element, dimensional element and reliability element:

$$M_n^t = (M_p^t + M_v^t) * M_d^t \tag{1}$$

The object is tracked for whole scene of 4K image and there is no need to track multiple times locally for each virtual camera view, this is done with aim to make computations more effective. The coordinates of object position, width w, and height h are calculated in real time using starting point of virtual view $[x_n, y_n]$, virtual view image ratio r_n and also height p_n of virtual view in pixels. The state of tracked object for n virtual view could by represented by vector:

$$O_n^t = [x - x_n, y - (y_n - p_n), w, h, d] \tag{2}$$

The dimensional element of metric is calculated from ratio of width of tracked object and width of virtual view:

$$M_v^t = \frac{w}{p_n \times r_n} \tag{3}$$

The reliability element is composed of the percentage expression of reliability of detection of the object being tracked using TLD algorithm. In general, the same number is created in specific time for each virtual view from overall scene. The positional element of metric is calculated from zonal division (Fig. 4) of each virtual view.

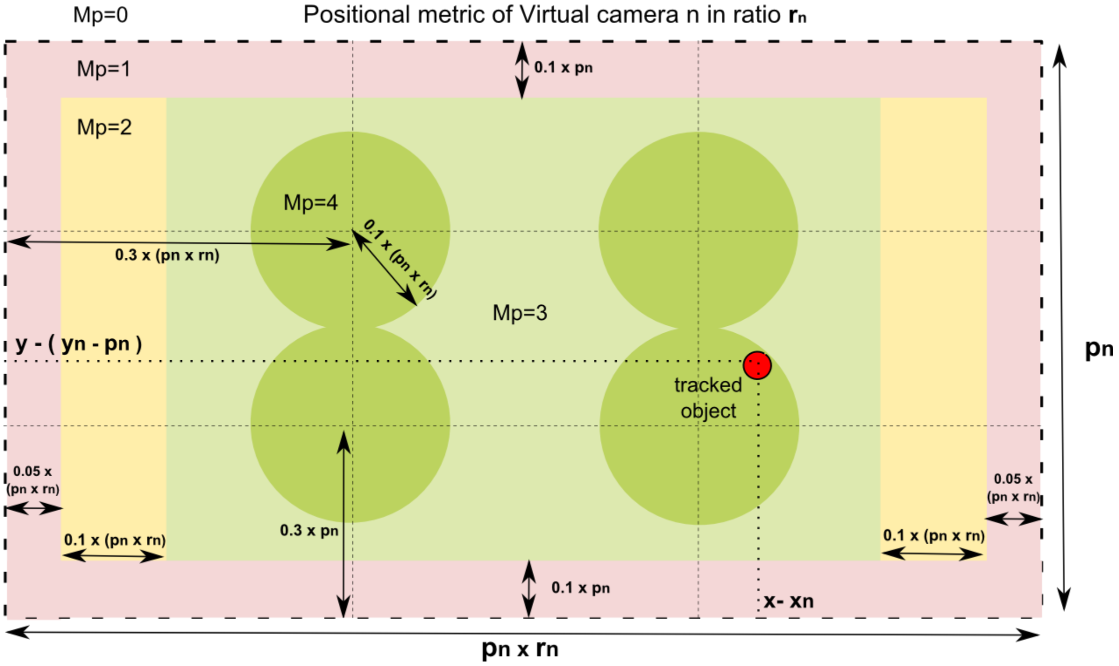

Fig. 4. Zonal division for calculating the positional metric for virtual view.

The zones are divided related to optimal composition. The rate is higher related how far the object is from intersections of thirds of virtual view. Revealing if the object is in near to optimal composition is based on several conditions with inequalities (4):

$$\begin{gathered}\left(x-x_n-\frac{p_n\times r_n}{3}\right)^2+\left(y-(y_n-p_n)-\frac{p_n}{3}\right)^2<\left(\frac{p_n\times r_n}{10}\right)^2\\ \left(x-x_n-\frac{2\times p_n\times r_n}{3}\right)^2+\left(y-(y_n-p_n)-\frac{p_n}{3}\right)^2<\left(\frac{p_n\times r_n}{10}\right)^2\\ \left(x-x_n-\frac{p_n\times r_n}{3}\right)^2+\left(y-(y_n-p_n)-\frac{2\times p_n}{3}\right)^2<\left(\frac{p_n\times r_n}{10}\right)^2\\ \left(x-x_n-\frac{2\times p_n\times r_n}{3}\right)^2+\left(y-(y_n-p_n)-\frac{2\times p_n}{3}\right)^2<\left(\frac{p_n\times r_n}{10}\right)^2\end{gathered} \tag{4}$$

The overall metrics in time t for each virtual view is stored to the evaluation vector: $H_t = [M_1^t, M_2^t, \dots M_n^t]$. Calculating the formula $\max_{0<k\leq n} M_k^t$ of all metrics in real time will give the source with the best metric in real time. This source has the best view on tracked object and will be switched to the output of live stream by cropping the overall 4K image by specific virtual view dimensions. Enhanced experience of switching the virtual view with object's best appearance without latency could be calculated by using k historic values of object parameters for each virtual view to predict the state of the object in one video frame forward (5).

$$PO_k = \begin{pmatrix} x^t - x_n & y^t-(y_n-p_n) & w^t & h^t & d^t \\ x^{t-1}-x_n & y^{t-1}-(y_n-p_n) & w^{t-1} & h^{t-1} & d^{t-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x^{t-k}-x_n & y^{t-k}-(y_n-p_n) & w^{t-k} & h^{t-k} & d^{t-k} \end{pmatrix} \tag{5}$$

4. Experimental Realization of Proposed Model

System based on interconnecting the vMix system with 2x SDI 4K video capture card with RED 4K camera was implemented. One capture card is handling the tracking of object using the TLD algorithm. Another is used to crop the image in vMix system. This test environment allows us to create multiple various sized virtual views that are automatically switched to the output with FFmpeg encoder based on calculated evolutional metric in real time based on object tracking results from TLD (Fig. 5).

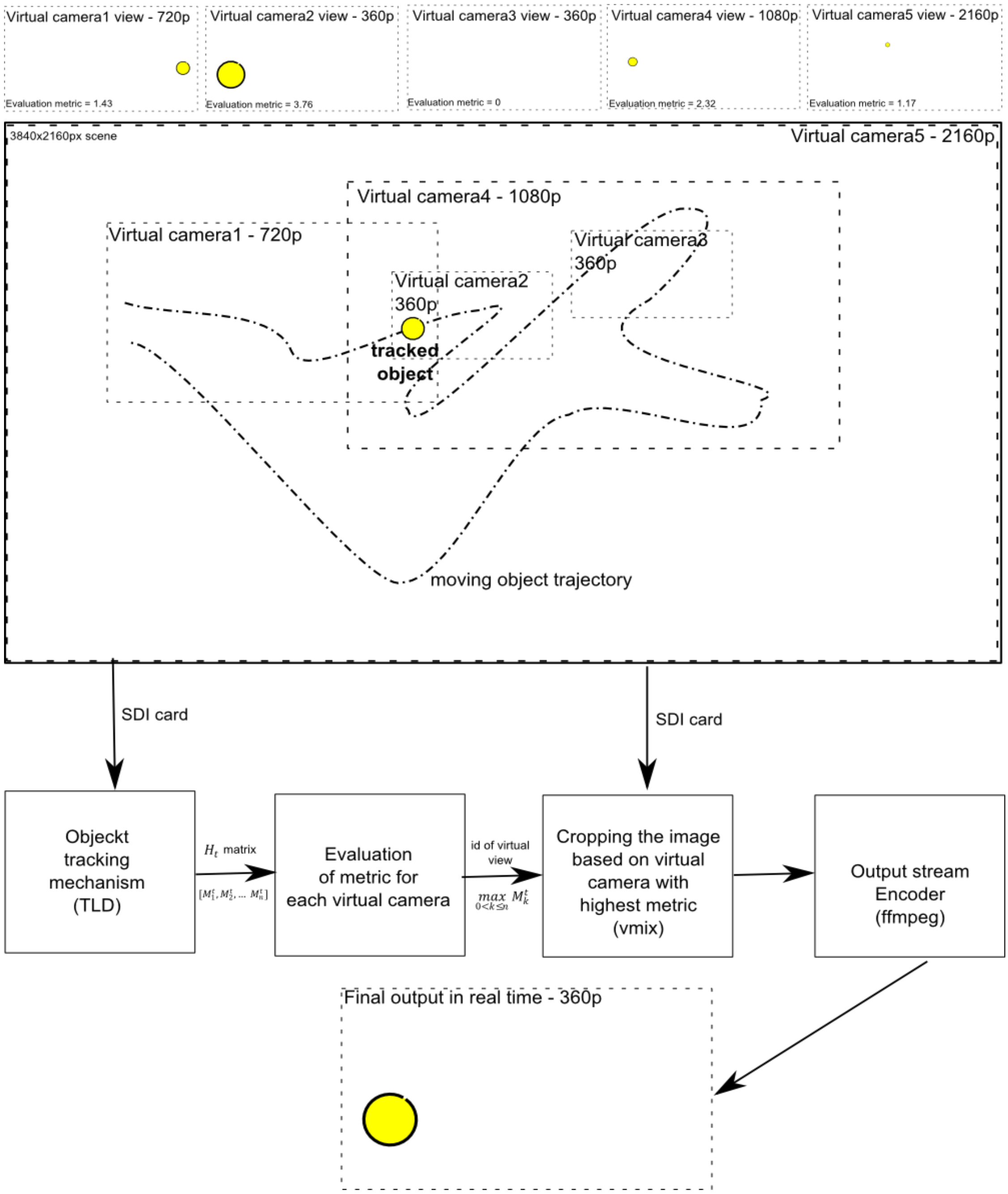

Fig. 5. Interconnection between components of experimental proposed solution.

5. Conclusion

In this study, we presented a new approach for automatic selection of the best shot on the object of interest in a real-time utilizing 4K definition-capturing device. Tracking of the object is utilized using TLD algorithm while selecting the best shot from the 4K video with virtual cropped view. This selection is automatic and based on the pre-learned pattern. The proposed model was implemented and tested in the laboratory environment. As for the future work the plan test the system in the real world conditions.

Acknowledgment

We support research activities in Slovakia/this project is being co-financed by the European Union. Paper is the result of the Project implementation: University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182, supported by the Research & Development Operational Programme funded by the ERDF.

References

[1] Wu, Y., Lim, J., & Yang, M.-H. (2013). Online object tracking: A benchmark. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2411-2418).

[2] Jarboe, G. (2012). *YouTube and Video Marketing: An Hour a Day* (2nd ed.). John Wiley & Sons, Inc.

[3] Shen, Y., Wu, L., Li, Y., Liu, S., & Wen, Q. (2014). Parallel data transfer protocol. *Enabling the New Era of Cloud Computing: Data Security*, *Transfer, and Management*, 238-255.

[4] Fujii, T., Kitamura, M., Murooka, T., Shirai, D., Fujii, T., & Takahara, A. (2009). 4K & 2K multi-resolution video communication with 60 fps over IP networks using JPEG2000. *Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems* (pp. 167-170).

[5] Seeling, P., & Reisslein, M. (2014). I Want Pixels (Entering the age of 4k). *IEEE Potentials*, *33(6)*, 27-30.

[6] Kalal, Z. (2010). Tracking learning detection. Retrieved December 18, 2014, from http://cmp.felk.cvut.cz/~matas/papers/kalal-2010-phd.pdf

[7] Kalal, Z., Mikolajczyk, K., & Matas, J. (2010). Face-TLD: Tracking-learning-detection applied to faces. *Proceedings of 17th IEEE International Conference on Image Processing* (pp. 3789-3792).

[8] Shi, H., Lin, Z., Tang, W., Liao, B., Wang, J., & Zheng, L. (2014). A robust hand tracking approach based on modified tracking-learning-detection algorithm. *Multimedia and Ubiquitous Engineering*, *308*, 9-15.

[9] Piao, S., & Berns, K. (2014). Multi-object tracking based on tracking-learning-detection framework. *Field and Assistive Robotics — Advances in Systems and Algorithms*, 74-87.

[10] Nebehay, G. (2013). Robust object tracking based on tracking-learning-detection. Retrieved December 23, 2014, from http://gnebehay.github.io/OpenTLD/gnebehay_thesis_msc.pdf

[11] Cymbalák, D., Kainz, O., & Jakab, F. (2014). Extended object tracking and stream control model based on predictive evaluation metric of multiple-angled streams. *International Journal of Computer Theory and Engineering*, *7(5)*, 343-348.

David Cymbalak received his M.Sc. and Ph.D. degrees in informatics from Technical University in Kosice, Slovakia. For more than 4 years, he has been a member of the well-recognized research institution — Computer Networks Laboratory at the Department of Computers and Informatics, Technical University of Kosice (DCI TUKE). His research includes multimedia content delivery, video streaming services and computer vision. He is also an active member of Slovakia startup community and co-founder of couple of them.

Ondrej Kainz was born in 1988. In 2013, he graduated with the M.Sc. dgree in applied informatics from the Technical University, Kosice, Slovakia. Since the very same year, he became a Ph.D. student at the Department of Computers and Informatics of the Faculty of Electrical Engineering and Informatics, Technical University of Kosice. His scientific research interests include image processing, computer vision, computer graphics, computer networks, e-learning and human-computer interfaces.

František Jakab was born in 1959. He received his M.Sc. degree in system engineering from St. Petersburg Electro Technical University, Russia, his Ph.D. degree in informatics from Technical University of Kosice, Slovakia. He is an associated professor at Technical University of Kosice where he established the well-known Computer Networks Laboratory (www.cnl.sk, in 1995). He has extensive experience in networking and utilization of ICT in education. He has been a coordinator of several large international projects financed by EC, coordinator of national wide ICT projects and research grants. He acts as an expert in areas such as projecting of computer networks, new form of multimedia based on communication (videoconferences, IP streaming, IPTV). He is an IEEE member and published more than 150 papers, 5 books; a chair of international symposiums and conferences, an editor of conference proceedings. He is the head of the Application Section of the Communication Technology Forum Association, head of Committee on Business — Academic Cooperation, American Chamber of Commerce in the Slovak Republic and the general manager of University Centre for Innovation, Technology Transfer and Intellectual Property Protection at Technical University of Kosice. He was awarded as the "IT person of the year 2006" in Slovak Republic.