ACTIVITY-BASED SYNTHESIZED FRAME GENERATION IN 3DTI VIDEO

Chien-nan Chen and Klara Nahrstedt
Department of Computer Science, University of Illinois at Urbana-Champaign
{cchen116, klara}@illinois.edu

ABSTRACT

In view of the high resource demand of 3D Tele-immersion (3DTI), we propose a user activity-based resource adaptation scheme that adjusts the compression ratio of 3DTI videos according to the user activity. The compression technique we designed is based on the frame synthesis via feature-based morphing. This technique is customized for 3DTI video due to the special properties of its scenes and the depth information provided by the system. Via a machine learning approach, our system can classify the user's current activity and choose the most suitable priority among temporal resolution, spatial resolution, and resource consumption in the adaptation. Our results show that the resource saving can reach up to 25% without perceptible degradation of the 3DTI videos.

Index Terms— *3D tele-immersion, QoE, morphing, resource adaptation, SVM*

1. INTRODUCTION

Effective remote interaction with media-rich visual content among geographically separated parties has come to its realization with the support of high-speed networking availability. Over the last decade, we have witnessed the thriving of 2D video conferencing. Currently, the interest of both academia and industry has moved on towards the development of multi-user 3D interaction. The 3D Tele-Immersion (3DTI) technology allows fully-body, multi-modal interaction among users, which opens a variety of possibilities on cyber collaboration and communication including e-learning [1], remote therapy [2], collaborative art performance [3], and interactive gaming [4].

Nevertheless, with its great potential, the quality of service (QoS) demands of 3DTI rise inevitably due to the real-time constraint of interaction, the processing complexity of graphic rendering, and the heavy load of data transmission [5]. These obstacles are challenging researchers to come up with more efficient utilization and allocation of the limited computing and networking resources. From a system-centric view, some scheduling [5] and synchronization [6] optimizations were proposed to improve the QoS in 3DTI. Yet, the complexity of the overall system proposed renders the application more difficult to be deployed.

Echoing the trend of human-centric computing, this work takes a different approach and exploits the sensibility of users under different usage scenarios of 3DTI system to achieve resource saving without perceptible changes in quality of experience (QoE). While most commercial 3D systems are specialized for sole purpose (e.g., Kinect being optimized for gaming), the 3DTI system is a platform designed for a wide range of user activities from low motion e-learning lectures to high motion action games. Based on previous observation [7], participants can have very different tolerance to quality degradation in 3DTI videos when viewing different types of user activities.

Taking advantage of the unique properties of 3DTI scenes and the depth information captured by the 3D camera, this paper proposes a frame rate boosting technique via feature-based morphing. The technique allows our system to perform tradeoffs between resource usage, temporal resolution, and spatial resolution of frames. Thus, its extended quality metric, the SFPS (Synthesized Frames per Second), serves two purposes, quality enhancement under the same resource limit; and resource saving without comparable QoE degradation.

Combining the two components, we propose a user activity-based resource adaptation scheme for 3DTI. The scheme connects the motion sensor data, the user activity, and the QoS demands via machine learning and subjective experiment. As a result, our proposed system of 3DTI videos is able to adjust in real-time its resource consumption according to resolution demands of different user activities. With the proposed scheme, the amount of resource saved reaches 25% for certain user activities with imperceptible degradation.

The contribution of this work is four-folds. 1) Proposition and evaluation of the application of morphing technique in 3DTI video compression and enhancement. 2) Proposition of the SFPS metric which bridges the tradeoff between resource consumption, temporal and spatial resolution in 3DTI video production. 3) Investigation of the relationship between the motion characteristics of common user activities in 3DTI environment and the noticeability of degradation on temporal/spatial resolution of the viewers. 4) Construction of a human-centric resource adaptation scheme that saves fair portion of computing and networking resource without comparable degradation of QoE.

The remainder of this paper is organized as follows. In the next section, an overview of previous works on resource adaptation in 3DTI is provided. Followed by more detailed

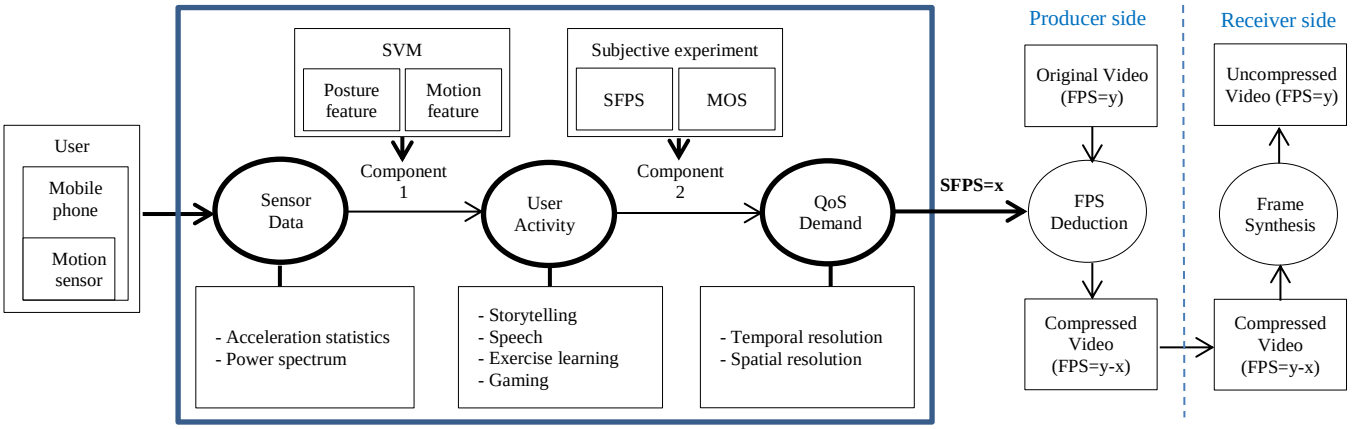

Fig. 1. Overview of user activity-based resource adaptation for 3DTI videos.

descriptions, section 3 shows the roadmap towards our resource adaptation scheme. As the first component, the machine learning based user activity classification is detailed in section 4. Section 5 and 6 introduce the second component, which includes the morphing technique for frame synthesis and the investigation of its relationship with QoE. In section 7, the results are evaluated and combined into the adaptation scheme proposed. Finally in section 8 we conclude.

2. RELATED WORKS

In this section, we review some adaptation schemes and frameworks in 3DTI. By their different design agendas and techniques, we classify them into adaptation based on application-layer semantics, human perception, and resource-performance balancing.

Adaptation Based on Application Layer Semantics. In earlier works, design of adaptation schemes tend to focus on per-stream basis, meaning, correlation and differentiation among streams are ignored. This implies waste of resource on delivering less important data to the application layer. Thus, [5] and [20] introduces application layer semantics to the design of adaptation scheme in 3DTI in order to achieve efficient dissemination. However, this introduces extra complexity to data sharing in the overlay network. Because every link in the overlay network is transmitting different parts of the stream bundle [19], a combination of streams containing different sensing data that are highly correlated, the topology of the network becomes a crucial factor that decides the efficiency of content dissemination.

Adaptation Based on Human Perception. The effective end-to-end transport of delay-sensitive data has long been a subject of study in interactive multimedia services. The goal of [8] is to provide human-centric adaptation on media playout scheduling (MPS) in 3DTI. The authors investigate the mappings between CMOS [9] and networking metrics to find the suitable adaptation for gross-motor and fine-motion user activities. Different from our purpose, the work focuses on resource allocation among streams rather than overall reduction of resource consumption.

Adaptation Based on Resource-Performance Balancing. There exist limits upon human perceptions, especially towards multimedia services. As a human nature, users are not keen enough to tell the differences between certain levels of service quality. Two QoE thresholds: Just Noticeable Degradation (JNDG) and Just Unacceptable Degradation (JUADG) were identified in [7] on the color-plus-depth level-of-details metric of 3DTI model rendering. With the two thresholds, the authors are able to adapt the resource consumption without devastating the service quality.

3. OVERALL APPROACH

Our total solution of activity-based resource adaptation can be broken down into two components (Fig. 1). The first component maps between motion sensor data and user activities; while the second component maps between user activities and QoS demand. With the two mappings, our adaptation can choose the most suitable setting for the compression of the video.

4. FROM SENSOR DATA TO USER ACTIVITIES

As the first component of our system, we built the mapping from the data acquired by motion sensors embedded in user's mobile phone to user activities with different motion characteristics.

4.1. User activity

Common user activities in 3DTI environment include e-lectures, exercise learning [1], and action gaming [4]. Each of the activities has its motional and postural uniqueness. Thus, by monitoring the readings of accelerometer attached to user, the activities can be classified in real-time with a machine leaning approach.

Fig. 2. An example of FPS boosting.

The activities we are targeting and their motional/postural characteristics are listed as follows, ordered by their degree of motion from low to high.

- Storytelling: user is sitting in the center of the 3DTI environment with most of his/her action concentrate on facial area. Occasional hand movement (e.g., page turning) is expected.
- Speech: user is standing in the center of the 3DTI environment. Frequent facial movement is expected along with occasional gesture and body movement.
- Exercise learning: user is mimicking the moves of a remote trainer or a physiotherapist. Slow and gross-motor movements of all body parts are expected.
- Gaming: both posture and position of the user in the 3DTI environment are changing rapidly. Fast and gross-motor movements are expected at all times.

4.2. Sensor data

By monitoring the orientation of the on-body accelerometer, posture of the user can be inferred. In order to minimize the complexity of usage, we do not require users to wear extra sensing devices on their body parts. Instead, we directly utilize the tri-axial accelerometer embedded in the user-owned mobile phones.

Based on the sizes of modern smart phones that are equipped with motion sensors, we assume that users are most likely to put them in their pants pocket when being asked to carry their phones on-body. Thus, the orientation of the device becomes distinctly different in sitting and standing scenarios and hence affects the proportion of the average accelerations of the three axes.

In addition, different speeds and changing frequencies of movements have direct effect on the variation of acceleration in the time domain and the power spectrum in the frequency domain. The former can be acquired by a sliding window analysis on the data compiled while the latter can be acquired by Fast Fourier Transform.

4.3. Activity classification via SVM

To build up a mapping from sensor data to user activity for real-time classification, we apply the Support Vector Machine [10]. The features, (e.g., positions, speed, changing frequency), fed into the machine are the aforementioned ones in the previous subsection and some variations (e.g., different window sizes for variance and average calculation.) The data were collected with four participants wearing ten different lower-body clothing (e.g., pants, jeans, basketball shorts.) Each participant was required to perform the four user activities with a Galaxy Nexus 4 Android phone [11] recording their acceleration in their pants pockets for five minutes. In sum, the total length of recorded data is 200 minutes, containing four targeted user activities with 50 minutes data each. With this training data, the SVM is able to generate a user activity classifier for our system which takes real-time sensor data as input, and output the type of user activity.

5. FPS BOOSTING

In this section we introduce the quality metric we are going to exploit: the Synthesized Frames per Second (SFPS). Due to the fact that we are targeting resource adaptation for user activities with different degrees of motion, naturally the prominence of this quality metric has to depend heavily on the motion of the video content.

For regular 2D videos, an intuitive choice would be frames per second (FPS) since its effect on QoE is closely correlated with the temporal characteristic of content [12]. Nevertheless, the range of adaptation of FPS for 2D videos is bounded by the limit of hardware (e.g., sampling rate of camera.) Thus, under common scenarios, most 2D video players choose to display the frames by its best effort with limited adaptation.

3DTI videos, on the other hand, can have more flexibility on adaptation of frame rate with the frame synthesis technique we propose. 3DTI videos bear some properties that make our synthesis possible:

- With the depth information provided by the 3D camera, only the subjects are captured to build 3D models to be put into the virtual space. The background of the scene is discarded.
- Common subjects in 3DTI are human bodies, which have fair sizes that take major portion of the scene.
- The number of subjects in one scene is restricted due to the interactive characteristic of the application and the size of the display. In most user activities, the number of subjects is expected to be no more than three, which lowers the graphical complexity of each scene.

Combining these properties, we propose to create synthesized frames by graphic morphing [13], which is a special effect in motion picture that transit one image to another based on predefined feature pairs. Figure 2 shows an example of applying morphing technique for frame synthesis. The first and the last frames in the figure are the only real frames captured by a camera, while the ten frames in between are all generated by morphing technique and hence boost up the original frame rate by eleven times. The morphing technique is only suitable for 3DTI videos but not

Fig. 3. An example of feature matching.

regular 2D ones because it can cause distortion on the background. The whole scene will be distorted along the morphing line pairs [13] marked on the subject if the subject is not separated from the background beforehand.

The state-of-the-art 3DTI system produces 3D videos with around 10 FPS [7]. This generates thousands of real frames in a clip in minutes. Thus, the marking of the morphing line pairs must be automatically done. We use feature detection [14] and feature matching [15] to mark the different location of the same feature in two frames (Fig. 3.) Due to the limited number of subjects in the scene of 3DTI videos and the fair size of the subjects, this feature-based matching can provide a meaningful number of matching pairs to the next phase. Given the matching features, a planar graph is built using Delaunay triangulation [16] on each frame with the detected features as vertices (Fig. 4.) The edges connecting the same pair of features become the line pairs for morphing. For image processing after the line pairs are acquired, the author encourage interested readers to refer to the paper of Beier and Neely [13].

As a result, with the morphing technique, we can either insert extra synthesized frames to boost up the original frame rate or we can replace some of the real frames with synthesized frames to save resource. SFPS affects both temporal and spatial resolutions of the final video. On the temporal side, adding extra synthesized frames can boost up the frame rate; while on the spatial side, because the synthesized frames have inferior graphical quality comparing to the real frames (due to possible feature mismatch), when we increase the proportion of synthesized frames (SFPS/FPS), the average spatial quality of all frames is compromised. Therefore, when adjusting the SFPS, we are actually arranging the priority of spatial resolution, temporal resolution, and resource consumption. For different video contents, the optimum tradeoff strategies can be very different. In the next section, we will investigate this complication.

6. FROM ACTIVITIES TO QUALITY DEMANDS

The need of resource adaptation for 3DTI applications comes from their high QoS demands. Therefore, in this section, the investigation of QoS demand on different user activities will be focusing on resource saving. We fix the temporal resolution (FPS) and concentrate on the tradeoff between spatial resolution and the compression ratio of videos. In practice, the compression is done by decreasing

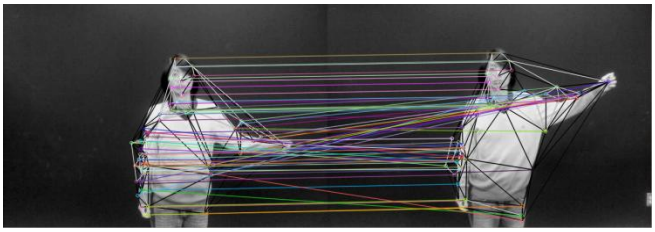

Fig. 4. An example of line pairs marking.

Speech | Storytelling | Exercise learning | Gaming

Fig. 5. 3DTI videos of the four user activities.

the FPS on the producer side (Fig. 1.) For a tolerable SFPS=x and the original FPS=y, the producer can decrease its FPS from y to y-x. On the receiver side, the deducted x frames per second can be recovered by morphing technique, restoring the FPS back to y with synthesized frames. Thus, both production and transmission resources are saved, and the compression ratio is 1-x/y. Our intention in this section is to find the maximum x (minimum compression ratio) without perceptible degradation for each user activity. (Note that x may not be an integer. For example, if we replace one real frame by synthesized one every two seconds then x=0.5.)

For each of the four user activities targeted, a 30 seconds clip is made. The subject in all videos is the same 26-year-old male with a 5'7" height. For speech and storytelling scenarios, the subject is speaking to the camera in his standing/sitting position. For exercise learning, the subject is learning Tai-Chi exercise in slow motion. For gaming scenario, the subject is participating in a lightsaber fencing match with another remote party (Fig. 5.)

The original clips are produced in y=10 FPS. These clips are referred as the reference clips in the following discussion. To mimic the video viewed in the receiver side, different numbers of real frames in the reference clips are replaced by synthesized ones. In total, 16 clips are generated for the four user activities with four different SFPS: 0.0, 1.0, 1.4, and 2.5. Respectively, the compression ratios of these test clips are 1.00, 0.90, 0.86, and 0.75.

15 participants (5 females, 10 males) are recruited to view and give scores to the visual quality of the clips on scale of 1 to 5 (Mean Opinion Score [17].) The maximum, average, and minimum ages of the participants are 50, 26.5, and 21, respectively.

To calibrate biased scores due to fatigue, we take the Absolute Category Rating with Hidden Reference (ACR-HR) [9] approach. The main idea of ACR-HR is to play a reference clip before each test clip so that the participants can adjust the score of each test clip based on the

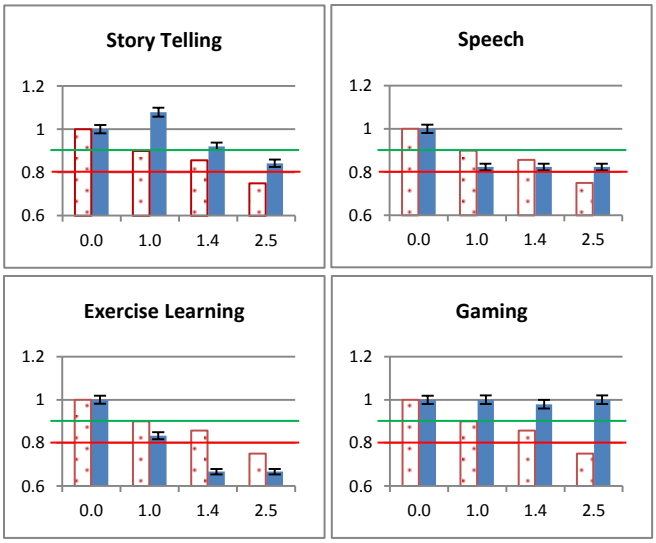

Fig. 6. Normalized MOS (solid blue) and compression ratio (dotted red) under different SFPS (x-axis).

corresponding reference. In addition, this arrangement is kept secret from the participants to minimize the effect of anticipation. In result, the length of the experiment sequence for one user activity with four test clips doubles (240 seconds), which makes the total time of the experiment approximately 20 minutes.

7. EXPERIMENT RESULTS AND ANALYSIS

The results of our experiments are presented and analyzed in two parts, which correspond to the user activity classification component and the quality demand component. After that, the results given by the two components are combined so as to become the total solution of our activity-based resource adaptation scheme.

7.1. User activity classification

The accuracies of classification on each user activity are listed in table 1. The evaluation of the classification is done by 10-fold cross-validation [10]. In result, the overall accuracy of classification is 91.5%.

Table 1. Accuracy of user activity classification

True\Inferred	Storytelling	Speech	Exercise learning	Gaming
Storytelling	100%	0.0%	0.0%	0.0%
Speech	0.0%	84.8%	14.4%	0.8%
Exercise learning	0.0%	7.6%	87.6%	4.8%
Gaming	0.0%	0.2%	6.2%	93.6

We can see from the table that speech and exercise learning scenarios have the lowest accuracy. The fact that the false positives of the two happen when users are actually performing the other activity implies our classification scheme is prone to confuse the two activities together. This is inevitable due to the observation that during exercise learning activity, the subject occasionally needs to stop his/her move and stands still to observe the demonstration of the trainer. According to the sensor data, this situation becomes identical to the speech activity.

Another observation is the perfect identification of the storytelling activity. This shows that the detection of posture signifies its uniqueness. However, note that this result is provided when the sensors are located in the users' pants pocket. To free this constraint, the perfect accuracy of sitting/standing classification should be sacrificed.

7.2. Quality demand

In figure 6, the normalized MOS (Mean Opinion Score) and the compression ratio of each test clip are plotted together to show the tradeoff between resolution and resource saving. For the solid blue bars with the 95% confidence interval marked, the y-axis stands for the average score given by the participants normalized by the score given to the reference clip[1]. For the dotted red bars, the y-axis is the compression ratio. The x-axis stands for SFPS.

In the storytelling activity, we can see that the synthesized frames blend in successfully. For SFPS=1.0, participants even give higher scores then the reference clip. This means that the degradation on spatial resolution caused by synthesized frames is imperceptible to the participants. In a storytelling scene, the limited movement of the subject's body minimizes the difference between frames, which makes the morphing result more authentic. In addition, the size of the subject is smaller due to the sitting posture, making the degradation even harder to be noticed.

The influence of motion on the prominence of spatial resolution can be seen in the comparison of speech and exercise learning scenarios. Again, the low motion (speech) of the video content lowers the bar of synthesis of an authentic frame. On the other hand, the gross body movements in the high motion scene (exercise learning) make the viewer concentrate more on the details of the subject's body, making the spatial degradation introduced by synthesized frames more detectable and hence bring down the MOS.

Following the same logic, the gaming scenario ought to be the most vulnerable one to SFPS degradation. However, we discover that the relationship between user's demand on spatial resolution and motion is not intuitive. In the contrary, the scores of the gaming scenario are the highest ones among all user activities. The cause of this phenomenon is revealed in the feedbacks of the participants. When asked about noticeable degradations in the gaming clips, participant W.C. (male, age 21) stated that "The unpredictable rapid moves of the subject make it hard to concentrate on the details", "All the jumping around and the

[1] The average scores of the reference clips are 4.3, 4.0, 4.2, and 4.0 for storytelling, speech, exercise learning, and gaming, respectively.

waving of the stick (lightsaber)… there are too many things going on in the scene." Even when the participants do notice the differences, they do not necessarily see them as drawbacks. Participant Z.G. (male, age 22) described the degradation as "really cool special effects". Z.G. further explained that "The motion blur of the lightsaber and the subject makes the fencing more exciting and enjoyable then the other clips." Apparently, after the degree of motion in the clip's content passes a certain threshold, its augmenting influence on the noticeability of SFPS degradation drops rapidly, making the tolerance of the degradation even higher than low motion activities.

7.3. Activity-based resource adaptation

In this final phase, we combine the aforementioned results together to build the adaptation scheme. Our purpose is to find the lowest resource consumption of the 3DTI videos without comparable degradation of the QoE. Thus, we setup two resource saving modes with different levels of QoE degradations: imperceptible (90% QoE) and acceptable (80% QoE.)[2] The two QoE thresholds are marked as green and red horizontal lines in figure 6, respectively. Table 2 shows the adjustment of SPFS in both modes with different user activities, and the compression ratio (CR) of each scenario. A demonstration of the compressed clips versus the references is provided as a supplementary material of this paper as well as on [18].

Table 2. Activity-based resource adaptation

		Storytelling	Speech	Exercise learning	Gaming
Impercept. mode	SFPS	1.4	0.0	0.0	2.5
	CR	0.86	1.00	1.00	0.75
Accept. mode	SFPS	2.5	2.5	1.0	2.5
	CR	0.75	0.75	0.90	0.75

8. CONCLUSION

In this work, we propose a morphing-based approach to synthesize frames in 3DTI videos. We further extend the technique to a quality metric: SFPS, which affects both temporal and spatial resolution of a video with different levels of resource consumption. We combine the adaptation of SFPS with motion characteristics of four common user activities in 3DTI environment. With a machine learning approach, the user activities can be classified in real-time based on the motion sensor data reported by users' mobile phones. Result shows that the degree of motion in the video content has significant influence on the prominence of SFPS's effect on QoE. However, the relationship is not intuitively monotonic. Finally, we build up a user activity-based resource adaptation scheme, which automatically classifies the user activity and assigns suitable SFPS to the production of the video. The scheme is able to save 25% of the rendering and networking resources without perceptible degradation on visual quality.

[2] With SFPS settings of 90% (80%) QoE requirement, more than 60% (20%) of the participants rate the compressed clips with equal or higher scores than the reference ones across all user activities.

9. REFERENCES

[1] R. Vasudevan et al., High quality visualization for geographically distributed 3D teleimmersive applications, TMM 2011.

[2] K. Nahrstedt, 3D Teleimmersion for remote injury assessment, International Workshop on Socially-Aware Multimedia, 2012

[3] M. Renata et al., Advancing interactive collaborative mediums through tele-immersive dance (TED): a symbiotic creativity and design environment for art and computer science, in Proc. of MM, 2008.

[4] W Wu et al., I'm the Jedi! - a case study of user experience in 3D tele-immersive gaming, in Proc. of ISM, 2010.

[5] Z. Yang et al., Enabling multi-party 3D tele-immersive environments with ViewCast, TOMCCAP, 2009.

[6] Z. Huang et al., SyncCast: synchronized dissemination in multi-site interactive 3D tele-immersion, in Proc. of MMSYS, 2011.

[7] W. Wu et al., Color-plus-depth level-of-detail in 3D tele-immersive video: a psychophysical approach, in Proc. of MM, 2011.

[8] Z. Huang et al., Perception-based playout scheduling for high-quality real-time interactive multimedia, in Proc. of INFOCOM, 2012.

[9] ITU-T P.910: Subjective video quality assessment methods for multimedia applications, 2008.

[10] C.-C. Chang et al., LIBSVM: a library for support vector machines, TIST, 2011.

[11] Nexus, http://www.google.com/nexus, 2012.

[12] L. Janowski et al., Content driven QoE assessment for video frame rate and frame resolution reduction, Multimedia Tools and Applications, Springer, 2011

[13] T. Beier et al., Feature-Based Image Metamorphosis, SIGGRAPH 1992.

[14] H. Bay et al., SURF: speeded up robust features, CVIU, 2008.

[15] M. Muja et al., Fast approximate nearest neighbors with automatic algorithm configuration, VISAPP , 2009.

[16] M. de Berg et al., Computational Geometry: Algorithms and Applications. Springer-Verlag. 2008.

[17] ITU-T P.10/G.100: Vocabulary for performance and quality of service, 2006.

[18] http://www.youtube.com/watch?v=ng-05WAc1zg

[19] P. Agarwal et al., Bundle of streams: concept and evaluation in distributed interactive multimedia environments, in Proc. of ISM, 2010.

[20] Z. Yang et al., A Multi-stream Adaptation Framework for Bandwidth Management in 3D Tele-immersion, in Proc. of NOSSDAV, 2006.