Video Selection for Visual Sensor Networks: a Motion-Based Ranking Algorithm

Simone Moretti, Matteo Mazzotti and Marco Chiani
DEI, University of Bologna, Italy
{simone.moretti6, mazzotti.matteo, marco.chiani}@unibo.it

***Abstract*—A Visual Sensor Network (VSN) is composed by several cameras, in general with different characteristics and orientations, which are used to cover a certain Area of Interest (AoI). To provide an optimal and autonomous exploitation of the VSN video streams, suitable algorithms are needed for selecting the cameras capable to guarantee the best video quality for the specific AoI in the scene. In this work, a novel content and context-aware camera ranking algorithm is proposed, with the goal to maximize the Quality of Experience (QoE) to the final user. The proposed algorithm takes into account the pose, camera resolution and frame rate, and the quantity of motion in the scene. Subjective tests are performed to compare the ranking of the algorithm with human ranking. Finally, the proposed ranking algorithm is compared with common objective video quality metrics and a previous ranking algorithm, confirming the validity of the approach.**

***Index Terms*—Visual Sensor Networks, QoE, camera selection techniques, ranking algorithms.**

I. Introduction

In recent years it has been widely acknowledged that the use of camera networks open the way to a large number of new applications. In particular, VSNs have been recently proposed in scenarios where a single camera could not provide a reliable coverage or a sufficient visual quality [1], [2]. Currently, VSNs are developed for surveillance of large areas, for environmental control of inaccessible or wild areas, and for telepresence in 3D remote video conferences. Furthermore, VSNs are a fundamental tool in tele-medicine to provide remote medical assistance and to support clinical therapy from a distance [3], [4]. One of the most critical issues is related to the large amount of data that each node collects from the monitored environment. In this regard, wireless camera networks require careful design and implementation of efficient radio resource allocation and node scheduling policies. The objective is to select, in each moment, the best camera subset that is able to satisfy some specific criteria: in literature this is defined as camera selection for large camera networks [1], [2], [5]. Several camera ranking proposals are based on geometrical considerations: the placement and orientation of the nodes are taken into account to define camera selection cost metrics. Each camera node is characterized by a directional sensing model. The gathered information depend on the direction on which the camera is oriented and the 3D viewing volume defined by the camera Field of View (FoV). The optimal camera deployment is also addressed in recent works [6], [7].

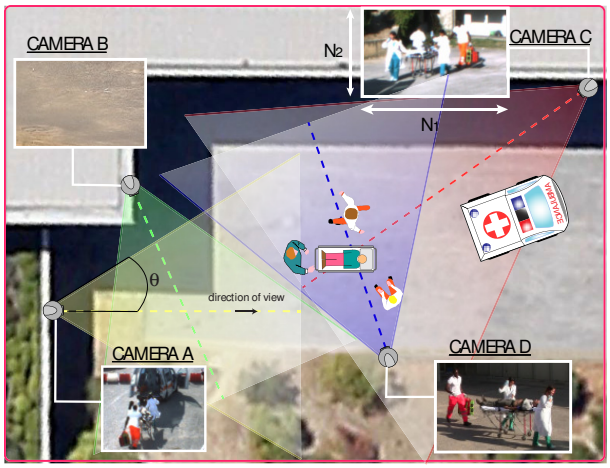

Fig. 1: Example of VSN

In this work, a novel content and context-aware camera ranking algorithm is proposed. Differently from the works in the literature, we approached the problem of the camera ranking with the aim to maximize the QoE to the final user. We designed the proposed ranking strategy taking into account the intrinsic parameters and resolution which characterize each camera. Furthermore, the camera selection is performed taking into account the camera position and related distance from the target of interest. Moreover, the proposed algorithm is based on the quantity of motion that is captured by each camera, and therefore denoted as Camera Ranking for Dynamic Scenes (CRDS) in the following.

II. Scenario Description

Let us consider the camera network scenario in Fig. 1, where the target (patient at the center of the figure) has to be monitored through the usage of a multi-camera system. Different wireless cameras with monitoring functionalities are placed in different positions. Each camera covers a particular portion of the considered area depending on its orientation and FoV. The knowledge of the point of interest allows a preliminary camera selection discarding all the cameras not capable to provide useful information for the final user. For example, camera B will be not taken into consideration in the final ranking. In this work, we assume that the cameras' position and attitude, are known in real-time, *e.g.*, by means of inertial units mounted on the devices. Modern equipment for emergency teams, in

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

fact, are nowadays more frequently capable to provide location information in order to improve the efficiency and the safety of the operations. Second, a pinhole model, which is the most widely used for mapping a 3D scene into a 2D image, has been adopted to describe the camera nodes. For a more realistic model, the pinhole is enriched with the camera FoV, which delimits the portion of the 3D monitored area. In Fig. 2-A, a 3D pinhole camera model has been depicted, with different vertical and horizontal FoVs. In the figure, the camera optical center is $\mathbf{c}$, the point to be monitored is $\mathbf{p}$, and $\mathbf{u}$ determines the camera orientation. Based on the camera pose and the point of interest, we delimit the AoI as illustrated in Fig. 2. The AoI is the 2D portion of the real-world in which the scene of interest is located. Considering the linear dependence between pixels and visualized area (in square meters), the proposed 3D camera model is decomposed in two 2D models where the number of linear pixels per meter is computed considering each FoV singularly. In addition, assuming the camera model has no rotations with respect to the global system reference, a geometric transformation is applied to translate the global system reference into the camera system reference. Thus, in Fig. 2-B, $\mathbf{p}' = \mathbf{p} - \mathbf{c}$ denotes the target position; $\mathbf{u}' = \mathbf{u} - \mathbf{c}$ describes the camera orientation; d is the distance between the optical center and the monitored target; finally, N_1 is the number of linear pixels along the first dimension of the image plane. The angle between the orientation vector $\mathbf{u}'$ and the target position $\mathbf{p}'$ is

$$\alpha = \arccos\left(\frac{\mathbf{p}' \cdot \mathbf{u}'}{|\mathbf{p}'| \cdot |\mathbf{u}'|}\right). \quad (1)$$

The projection of the distance $d = \|\mathbf{p}'\|$ on the optical axis is calculated as $d' = d\ \cos\alpha$, and the length visualized along the considered image plane is

$$M = 2\ d'\ \tan\theta\,. \quad (2)$$

III. Camera Ranking for Dynamic Scenes (CRDS)

In [5] a camera ranking algorithm, called Camera Ranking for Static Scenes (CRSS), has been proposed for low motion scenes. CRSS is based on the amount of pixels required to represent a unit area (1 square meter) at the target distance. In this section, a new algorithm for video with arbitrary motion called Camera Ranking for Dynamic Scenes (CRDS), is proposed.

Conceptually, the CRDS takes into account the spatial features considered in CRSS, but includes in the ranking criteria also the acquisition frame rate and the average motion feature velocity. In order to study the impact of frame rate and resolution on the final perceived video quality, we assume without loss of generality that each video sequence can be modelled as a three-dimensional continuous signal. The video signal is acquired by the cameras and then is discretized taking into account the frame rate and the spatial resolution as sampling frequencies. We start by defining a model for the intrinsic video power spectrum of the filmed scene. Then, the camera quality for the specific scene is estimated as the

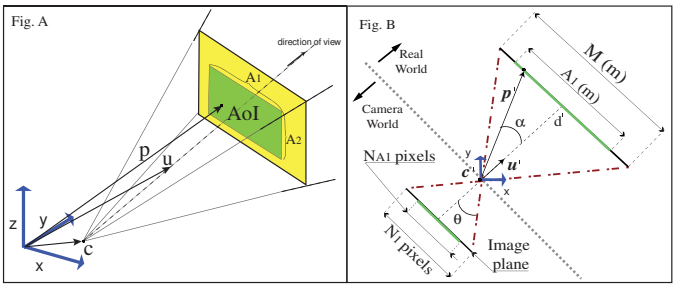

Fig. 2: The pinhole camera model: A) 3D representation; B) 2D representation.

amount of intrinsic video power that the camera can acquire. Given the sampling frequencies, this acquired power depends on the signal bandwidth in the spatial and temporal domains: the larger is the sampling frequency, the greater is the collected power by the signal. The spatial sampling rate depends on the geometry of the scenario and on the scaled resolution, based on the CRSS principle. The role of the temporal sampling rate depends on the scaled video motion. In this regard, we used the analysis leading to the concept of spatio-temporal power video spectrum, as proposed in [8]. In that work, the authors started by analyzing a thousands of video segments, in order to identify common regularities in natural scenes. A video segment is a 3-dimensional hypercube of size $L \times L \times T$, where the first two dimensions refer to the spatial coordinates of a point in a frame, while the third dimension identifies the time instant within the sequence. Indicating with $s(\mathbf{x}, t)$ the windowed light intensity in point $\mathbf{x} = (x_1, x_2)$ at time t, the correlation between two points separated by the spatio-temporal distance $(\boldsymbol{\xi}, \tau)$ can be expressed as:

$$r(\boldsymbol{\xi}, \tau) = \frac{1}{L^2 T} \int_0^L \int_0^L \int_0^T s(\mathbf{x} + \boldsymbol{\xi}, t + \tau) \cdot s(\mathbf{x}, t) dx_1 dx_2 dt \quad (3)$$

where the spatial displacement is $\boldsymbol{\xi} = (\xi_1, \xi_2)$.

The power spectrum of the considered spatio-temporal segment is obtained through the Fourier transform of $r(\boldsymbol{\xi}, \tau)$:

$$R(\mathbf{f}, w) = \int_{-L}^{L} \int_{-L}^{L} \int_{-T}^{T} r(\boldsymbol{\xi}, \tau) \cdot e^{j2\pi(\mathbf{f} \cdot \boldsymbol{\xi} + w\tau)} d\xi_1 d\xi_2 d\tau \quad (4)$$

where $\mathbf{f} = (f_1, f_2)$ are the spatial frequencies and w the temporal frequency, respectively. Assuming that the objects in the scene are placed at a distance in the range $[d_1, d_2]$ from the camera and that their static spectrum is rotationally symmetric, the following expression is derived for the average power spectrum [8]:

$$G(f, w) = \frac{K}{f^{m+1}} \int_{d_1}^{d_2} P\left(\frac{w}{f} z\right) z\, dz \quad (5)$$

where $f = ||\mathbf{f}||$, $P(\cdot)$ is the velocity distribution of the objects in the scene along a certain direction[1], and K and m are two parameters whose values are estimated numerically. Note that

[1]Note that in the model of [8] also the velocity distribution of the objects in the scene is assumed as rotationally invariant.

in (5) the power spectrum depends on the velocity distribution, which in many natural scenes can be approximated by simple power-law distribution [8]. Finally, the power spectrum of natural varying images is modeled as [8]:

$$G(f,w) = \frac{K\bar{v}}{2f^{m-1}w^2}\left[\frac{n-2}{(x+1)^{n-1}} - \frac{n-1}{(x+1)^{n-2}}\right]_{\frac{wd_1}{fv_0}}^{\frac{wd_2}{fv_0}} \quad (6)$$

where $\bar{v}$ is the average object velocity, and v_0, n are constant values. This analytical model was numerically validated in [8] by observing that the measured power spectrum of the considered video segments, once scaled by the factor f^{m+1} and expressed as a function of f/w, shows a behavior in total accordance to (6). Hence, the total collected power can be calculated as:

$$P = \int_0^{w_{ul}} \int_0^{f_{ul}} G(f,w)\, df\, dw. \quad (7)$$

For each camera, f_{ul} and w_{ul} depend on the spatial and temporal bandwidth for the considered video signal. According to the Shannon sampling theorem, the camera frame rate F_r acts as a temporal sampling factor transforming the continuous signal in a discretized version. Thus, the temporal bandwidth is $w_{ul} = F_r/2$.

To determine the spatial frequency, we should take into account (see Fig.2-B) the length visualized along the considered image plane *M*, given by (2), the distance between the camera optical center and the filmed scene *d'*, and the number of pixels forming the AoI on the image plane, N_{A1}. The spatial sampling frequency in cycles per degree is therefore

$$S_f = \arctan\left(\frac{M}{d' \cdot N_{A1}}\right)^{-1} [cycles/degree] \quad (8)$$

where $b = M/N_{A1}$ is the length projected on one linear pixel. The best spatial frequency is guaranteed when the AoI fullfill the whole image plane: more precisely, when the camera is placed at the distance d' such that that $N_1 = N_{A1}$ (Fig. 2-B). In this ideal case the scene of interest is sampled taking into account the whole spatial resolution of the considered camera. Increasing the distance, the projection of the AoI on the image plane gets smaller: respect to the ideal case, the spatial resolution N_{A1} is lower and the spatial sampling is coarser, resulting in lower spatial frequency. In accordance to the relation between spatial bandwidth and spatial frequency, the spatial bandwidth is $f_{ul} = S_f/2$.

Once the spatial and temporal bandwidth have been calculated for each camera, the CRDS uses (7) for ranking: the higher is the power, the higher will be the camera classification in the ranking.

IV. Experimental results

A. VSN parameters and AoI description

The CRDS algorithm has been experimentally validated on the videos of nine different cameras. In Tab. I we summarize the camera frame rate, the image dimensions, and the related AoI dimensions. The cameras are placed with the same pose respect to the object of interest, which is placed at a distance $d' = 5$ m. For simplicity, we first acquired the video from a high resolution and frame rate camera, C_1. Then, the video camera streams $C_2, \ldots, C_9$ have been virtually obtained by applying frame decimation and image resizing techniques on C_1. For example, the video sequence provided by C_6 is obtained applying spatial decimation factor equal to 2 and a temporal decimation factor equal to 4 on C_1. In Fig. 3 we provide a frame captured by C_1 and related AoI.

Camera (C_i)	Image dim.	AoI dim.	Frame rate (F_r)
1	720×576	320×320	25
2	720×576	320×320	12.5
3	720×576	320×320	6.25
4	360×288	160×160	25
5	360×288	160×160	12.5
6	360×288	160×160	6.25
7	180×144	80×80	25
8	180×144	80×80	12.5
9	180×144	80×80	6.25

TABLE I: Set of cameras used for testing the CRDS

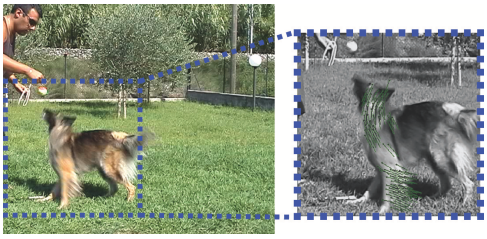

Fig. 3: Left: Frame extracted from C_1. Right: the related AoI.

B. Object of interest velocity extraction

In this section, we describe the feature velocity extraction technique we implemented to define the mean velocity of the object of interest within the AoI. In the first step, we calibrated camera C_1 (See Tab. I) through the Camera Calibration Toolbox for Matlab®. We extracted the optical flow using the *OpenCV* library [9]. In particular, we utilized a technique to find the dense optical flow based on Gunner Farneback's algorithm. For each i^{th} sub-block belonging to the AoI we calculated the related velocity in the following way:

$$v_i = \|m_i\|^2 \cdot q_{lin}^{-1} \cdot F_r \quad (9)$$

where $\|m_i\|^2$ is the displacement magnitude of the considered sub-block and $q_{lin} = N_1/M$ is the number of linear pixels used for describing a single meter at the considered distance *d'*= 5 m. In Fig. 3-right the displacement vectors in the AoI are indicated by the bold arrows. Then, to compute the average velocity $\bar{v}$, as requested in (6), we use the fact that, during a video visualization, the attention of the user is more focused on objects that are moving. In other words, the human-eye is more attracted from rapid object position variations, and less sensitive to what-is-happening in the background and in the contour features. Taking into account this behavior, the average velocity has been expurgated from all the features characterized by low or approximately null quantity of movement [9].

$C_{i,j}$	1	2	3	4	5	6	7	8	9
1		60	65	90	100	65	100	65	100
2	40		65	30	90	100	65	65	100
3	35	35		40	15	45	20	30	65
4	10	70	60		100	65	100	65	60
5	0	10	85	0		65	65	100	65
6	35	0	55	35	35		65	80	60
7	0	35	80	0	35	35		80	95
8	35	35	70	35	0	20	20		90
9	0	0	35	40	35	40	50	10	

TABLE II: PMOS results

With this approach, the average object of interest velocity for our experimental video sequences resulted to be $\bar{v} = 5.7$ m/s.

C. Subjective video quality assessment results

Subjective tests are typically used in the video processing field to obtain the human user's perception of the quality of the processed video sequences. In this work, the video quality assessment will be used as reference metric to rank the cameras based on a perceived quality point of view. We adopted a modified MOS version, namely Pairwise Mean Opinion Score (PMOS). In PMOS, the quality assessment is obtained visualising a pair of video sequences at each time. The viewer is asked to express his/her preference choosing the video sequence with the highest perceived quality. Each possible pairwise video combination is visualized and judged by the voters. We selected a population of 25 voters equally divided between males and females. The camera set is characterized by different spatial and temporal resolution and this makes a direct quality comparison of the AoI unfeasible. For this reason, the scene of interest is extracted based on the position of the point of interest and the camera position and orientation. Then, the video sequences focusing the scene of interest are re-sampled in the spatial and temporal domain to a common format, so that they can be directly compared through objective and subjective tests. In Tab. II the PMOS results have been proposed. Specifically, each element $C_{i,j}$ of Tab. II indicates the percentage of users which prefer the video sequence provided by C_i over C_j. For example, the 70% of voters prefers C_4 over C_2. In the final step, the Kemeny-Young method is applied to obtain a camera ranking based on the PMOS results. The Kemeny-Young technique has been developed to identify the most popular choices in an election exploiting a pairwise comparison and assigning a score to all possible ranking sequences. Each sequence considers which choice might be most popular, which choice might be second most popular, and so down to which choice might be least-popular. The ranking sequence obtaining the highest score is the selected one [10].

D. Camera ranking techniques comparison

The comparison between two different sorting metrics is a non-trivial problem: a great number of works has been proposed in literature with the aim at measuring the difference between sorting techniques. We implemented two known metrics, such as the Spearman's rank correlation coefficient and the Kendall's tau distance. These metrics calculate the possible correlation (or distance) between two different ranking strategies. In statistics, Spearman's Correlation Coefficient (SCC) is a non-parametric measure of statistical dependence between two variables [11]. The Kendall's Tau Distance (KTD) measures the total number of pairwise inversion between two ranking lists. The larger the distance, the more dissimilar the two lists are [12]. However, these metrics do not consider the ranking positions as well as the element relevance in the pairwise inversion counting: due to this reason, such metrics are called *invariant*. A general solution of this problem consists of weighting each ranking inversion by taking into account the element positional information or defining an element relevance criteria. Intuitively, a raking inversion on a high-weight element should be more penalizing than an inversion on a low-weight element. Following this approach, we defined a Weighted Kendall's Tau Distance (WKTD) version in which the element relevance is determined by the subjective test results. Let $[n] = 1, \ldots, n$ be a set of elements and W_n be the set of permutations on $[n]$. In its turn, $\psi \in W_n$ is the permutation provided by the PMOS results and $\psi(i)$ is the ranking of the i^{th} elements. Again, $\sigma \in W_n$ is the permutation provided by one the proposed camera ranking algorithms. The WKTD resulting distance between rankings, $\bar{K}(\psi, \sigma)$, is defined as:

$$\bar{K}(\psi, \sigma) = \frac{1}{\xi} \sum_{(i,j):i<j} K_{i,j}(\psi, \sigma) \tag{10}$$

where:

$$K_{i,j}(\psi, \sigma) = |w_{i,j}(\psi(i) < \psi(j) \wedge \sigma(i) > \sigma(j)) \vee w_{j,i}(\psi(i) > \psi(j) \wedge \sigma(i) < \sigma(j))| . \tag{11}$$

In the classical KTD formulation $w_{i,j} = 1$ and $w_{j,i} = 1$. In Tab. II, $C_{i,j}$ is the percentage of users who preferred element i over element j. Therefore, $C_{i,j}$ is also intended as the percentage of users who may be unsatisfied if element i and j are ranked in the opposite order. Hence, we propose to modify the KTD by setting the weights $w_{i,j} = C_{i,j}$ and $w_{j,i} = C_{j,i}$ in (10): the higher is the user preference on a certain ranking order, the more significant is the pairwise inversion counting in the considered metric. Accordingly, the normalization factor ξ is calculated in (12) taking into account all the possible pairwise inversions, as

$$\xi = \sum_{(i,j):i<j} |C_{i,j}(\psi(i) < \psi(j)) \vee C_{j,i}(\psi(i) > \psi(j))| . \tag{12}$$

V. Camera Ranking Validation

In Tab. III we collected the camera ranking results considering the objective metrics, the CRSS and the proposed CRDS. In particular, the position of camera C_i in each considered ranking techniques is provided. Furthermore, the subjective test results are also provided based on the Kemeny-Young method in the second column of Tab. III. We can notice the CRSS results mainly depend on the camera resolutions. This is due to the fact that the camera set is characterized by the same

C_i	PMOS	PSNR	SSIM	CRSS	CRDS
1	1	1	1	1	1
2	3	2	2	2	3
3	8	5	4	3	7
4	2	3	3	4	2
5	4	4	5	5	4
6	5	6	6	6	8
7	6	7	7	7	5
8	7	8	8	8	6
9	9	9	9	9	9

TABLE III: Subjective test results (second column) and camera ranking results for the implemented techniques.

	PSNR	SSIM	CRSS	CRDS
KTD	0.05	0.11	0.14	0.02
SCC	0.78	0.78	0.84	0.93
WKTD	0.09	0.14	0.19	0.02

TABLE IV: Distances and correlation between automatic and human ranking with $\bar{v} = 5.7$ m/s.

pose and this algorithm does not depends on the frame rate and the feature velocities. In its turn, several inversions occur if the ranking provided by the subjective tests are compared with the objective ones. In particular, we can observe how the objective metrics tend to favour high-resolution cameras; conversely, the subjective tests privilege cameras characterized by higher frame rate. The CRDS algorithm validation is composed by two steps. In the first step, we verified the average feature velocity computation and the feasibility of the assumption we made in Sec. IV-B. To this purpose, several CRDS results are provided varying the average feature velocity $\bar{v}$ in (6). In Fig. 4 these results are compared with the subjective tests using the KTD, WKTD and the complementary version of the Spearman's coefficient (C-SCC). We can notice that the application of CRDS using the average expurgated velocity we calculated in IV-B provides the best similarity between CRDS and subjective test results. In the second step, we provided a detailed evaluation of the proposed CRDS comparing the obtained results with the users' perceived video quality defined by the subjective tests. We can observe how the CRDS and the subjective test ranking results are identical for the top four cameras (*i.e.*, $C_{1,2,4,5}$) which can be identified as the best quality cameras. The main difference is related to the low-quality cameras which are $C_{6,7,8}$. Coherently, the proposed ranking algorithm tends to prefer higher frame rate cameras again, while the users' preference privilege lower frame rate cameras but higher resolution cameras in this case. In Tab. IV, the KTD, WKTD and SCC values are presented comparing the ranking based on subjective tests and the other proposed ranking techniques. We can notice that CRDS provides the best results, i.e., the lowest distances and the highest correlation, when compared with the human ranking.

VI. Conclusion

In this work, a novel camera ranking algorithm has been presented. The proposed algorithm is aimed to find the camera subset that best satisfies specific ranking criteria, and is designed to optimize the QoE for the final user in terms of video quality and significance. To validate the proposed technique, a multi-camera acquisition system has been arranged and the collected video sequences have been evaluated through objective and subjective tests. The experimental results show the effectiveness of the proposed ranking solution to provide the best user experience in terms of visual quality.

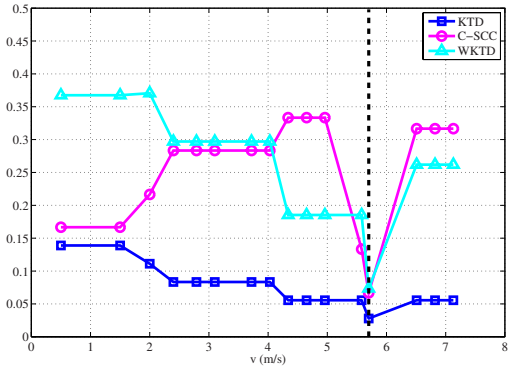

Fig. 4: Distances (KTD, C-SCC, and WKTD) between CRDS and the subjective tests, as a function of the parameter $\bar{v}$ in (6), for the recorded scene.

Acknowledgement

This work was supported in part by the European project EuroCPS (grant no. 644090) under the H2020 framework.

References

[1] S. Soro and W. Heinzelman, "Camera selection in visual sensor networks," in *Proc of AVSS '07*, 2007, pp. 81–86.
[2] A. Mavrinac and X. Chen, "Modeling coverage in camera networks: A survey," *Int. Jour. of Comp. Vision*, vol. 101, no. 1, pp. 205–226, 2013.
[3] S. Hengstler, D. Prashanth, S. Fong, and H. Aghajan, "Mesheye: a hybrid-resolution smart camera mote for applications in distributed intelligent surveillance," in *Proc. of IPSN 2007*, pp. 360–369.
[4] G. Eysenbach, "What is e-health?" *Journal of medical Internet research*, vol. 3, no. 2, 2001.
[5] S. Moretti, S. Cicalò, M. Mazzotti, V. Tralli, and M. Chiani, "Content/context-aware multiple camera selection and video adaptation for the support of m-health services," *Procedia Computer Science*, vol. 40, pp. 206–213, 2014.
[6] A. Ercan, D. Yang, A. E. Gamal, and L. Guibas, "Optimal placement and selection of camera network nodes for target localization," in *Distributed Computing in Sensor Systems*. Springer, 2006, pp. 389–404 vol. 4026.
[7] D. Yang, J. Shin, A. O. Ercan, and L. Guibas, "Sensor tasking for occupancy reasoning in a network of cameras," in *Proc. of the IEEE/ICST BASENETS '04*, San José, CA, USA, 2004.
[8] D. Dong and J. Atick, "Statistics of natural time-varying images," *Network: Comput. in Neural Syst.*, vol. 6, no. 3, pp. 345–358, 1995.
[9] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.
[10] H. P. Young and A. Levenglick, "A consistent extension of condorcets election principle," *SIAM Journal on Applied Mathematics*, vol. 35, no. 2, pp. 285–300, 1978.
[11] C. Spearman, "The proof and measurement of association between two things," *The Amer. Jour. of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.
[12] M. Kendall and J. Gibbons, *Rank correlation methods*, ser. A Charles Griffin Book. E. Arnold, 1990.