Leveraging Multiview Video Coding in Clustered Multimedia Sensor Networks

Stefania Colonnese*, Francesca Cuomo* and Tommaso Melodia†

* DIET, University of Rome, La Sapienza, Rome, Italy
† Department of Electrical Engineering, State University of New York at Buffalo, NY 14260, USA

***Abstract*—We experimentally characterize the compression efficiency of Multiview Video Coding (MVC) techniques in Wireless Multimedia Sensor network (WMSN) composed of multiple video cameras with possibly overlapping field of views. We derive an empirical model that predicts the compression efficiency as a function of the *common sensed area* (CSA) between different camera views. We show that the CSA depends not only on geometrical relationships among the relative positions of different cameras, but also on several *object-related phenomena*, e.g., occlusions and motion, and on *low-level phenomena* such as variations in illumination. We then apply the model to a WMSN, where we create clusters based on the CSA as estimated by exchanging local data. Based on this estimates, we form clusters and measure the resulting transmission rate. Numerical simulation results show that building clusters based on a CSA criterion can bring significant performance gains in terms of bandwidth efficiency. The herein presented promising results pave the way for clustering optimization taking into account different networks constraints and conditions.**

***Index Terms*—Video Sensor Networks, Multiview Video Coding, MVC Efficiency.**

I. Introduction

Wireless multimedia sensor networks (WMSNs) can support a broad variety of application-layer services [1], [2]. When different sensors cameras acquire different views of the same scene, multi-view oriented processing techniques enable tasks such as video scene summarization [3], moving object detection [4], face recognition [5], depth estimation for 3D video rendering [6], surveillance, and social robotics to cite a few.

Video coding techniques for multiview sequences (i.e., sequences in which the same video scene is captured from different perspectives) provide more compact video representations and thus enable efficient resource allocation. Cameras whose fields of view (FoV) are significantly overlapped can be jointly encoded through multiview video coding (MVC) techniques. Nevertheless, if MVC techniques are applied to sequences that differ significantly (for instance, because of the presence of different moving objects), MVC may provide equal or even lower compression performance than encoding each view independently. Hence, adoption of MVC may in some cases even be detrimental, whereas in other situations the efficiency of MVC compared to single view video coding (AVC) can be leveraged to optimally allocate resources at different layers of the protocol stack.

In this paper, after defining the notion of *common sensed area* (CSA) between different views, we experimentally characterize the relationship between MVC compression gain (with respect to single view video coding) and the estimated CSA between views and we show the benefits of the adoption of the CSA parameter as a clustering criterion in a WMSN.

The structure of the paper is as follows. In Section II, we discuss the multimedia sensor network model an review the state of the art in adopting MVC encoding in WMSNs. After introducing the notion of common sensed area in Section III, in Section IV we empirically establish the relationships between the efficiency of MVC and the common sensed area. In Section V we apply the developed empirical model to WMSN clustering. Finally, Section VI concludes the paper.

II. MVC in Multimedia Sensor Networks

A WMSN is typically composed of multiple cameras, with possibly overlapping FoVs. The FoV of a camera can be formally defined as a circular sector of extension dependent on the camera angular width, and oriented along the pointing direction of the camera. A given FoV typically encompasses static or moving objects at different depths positioned in front of a far-field still background. An illustrative example is reported in Figure 1.

The camera imaging device performs a radial projection of real-world object points into the camera plane where they are effectively acquired. The acquired image is usually referred to as the image plane. For instance, the image planes corresponding to the scenario in Fig. 1 are shown in Fig. 2.

While every point in the image plane has a corresponding point in the FoV, not all the points in the FoV correspond to points in the image plane, due to the occlusions between objects at different depths. We observe that while the FoVs depend on characteristics exclusively of the camera such as position, orientation, angular view depth, the effectively acquired images resulting from the projection of real-world objects on the image plane depend on the effectively observed scene. First, each near-field object partially occludes the effective camera view to an extent depending on the object size and on the object-to-camera distance. Besides, the same real-world object may be seen from different points of view and at different depths by different cameras. Therefore, the views provided by the nodes of a WMSN may correspond to image planes characterized by different degrees of similarity, depending both on the camera locations and on the framed

978-1-4673-0921-9/12/$31.00 ©2012 IEEE

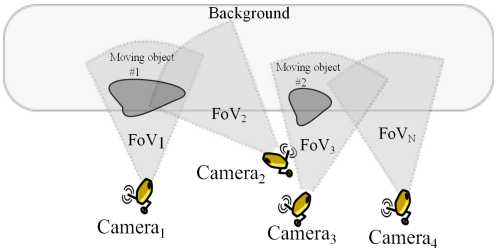


Fig. 1. Example scenario.

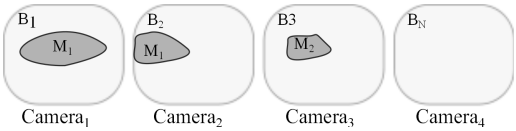


Fig. 2. Different image planes.

scene. The view similarity can be exploited to improve the compression efficiency through MVC.

The problem of compressing correlated data for transmission in WMSNs has been debated in recent papers. In [7], highly correlated sensors covering the same object of interest are paired to cooperatively perform the sensing task by capturing part of the image each. The pioneering work of [1] demonstrated that a camera geometry-based algorithm can be designed for selecting a suitable group of cameras communicating toward a sink so that the amount of information from the selected cameras can be maximized. A clustering scheme taking into account the angular cameras disparity is introduced in [8], such that the set of coding clusters covers the entire network with maximum compression ratio.

Nonetheless, even for cameras with low difference in the sensing direction, occlusions among real world foreground objects in the FoVs may cause the acquired images to differ significantly. Besides, motion of objects may result in time-varying inter-view similarity even if the WMSN nodes maintain the same relative positions. These observations motivate us to consider a different, scene-related, parameter accounting for key phenomena that affect the correlation between views, based on which model the efficiency of MVC techniques.

III. Common Sensed Area

In this Section, we introduce the notion of *common sensed area*. The latter depends not only on geometrical relationships among the relative positions of different cameras, but also on several real *object related phenomena*, namely occlusions, motion, and on *low-level phenomena* such as illumination changes. To this aim, we start by formulating a continuous model of the images acquired by different cameras.

The acquisition geometry is given by a set of N cameras, with assigned angular width and FoV (as in Fig. 1). Real-world objects framed by the video cameras are mapped into the image plane. Let us consider the luminance image $I^{(i)}(x,y)$ acquired at the i-th camera at a given time[1]. Each image point (x_i, y_i), represents the radial projection, on the i-th image plane, of a real point (u,v,z).

Let us now consider a pair of images $I^{(i)}(x,y), I^{(j)}(x,y)$, acquired by cameras with possibly overlapping FoVs. We first define the common area $\mathcal{C}$ between the two images i and j as the set of points (x_i, y_i) in the i-th image representing real points (u,v,z) appearing also in the j-th image, namely

$$\mathcal{C} = \Big\{ (x_i,y_i) \in I^{(i)}(x,y), (x_i,y_i) s.t. \\ \text{if } (x_i,y_i) = P_i(u,v,z), \text{then } \exists\ (x_j,y_j) = P_j(u,v,z) \in I^{(j)}(x,y) \Big\}$$

Although originated by the same object, the luminance values associated to (u,v,z) in images from different cameras differ because of the different perspective warping under which the scene is acquired and of several other acquisition factors, including noise and illumination.

Based on the afore defined sets, we can formally define the CSA between views. Let us consider the image $I^{(i)}[m,n]$ obtained by sampling $I^{(i)}(x,y)$ on a discrete grid with sampling interval $(\Delta x, \Delta y)$. We define the CSA $\alpha(i,j)$ as

$$\alpha(i,j) = \frac{|\mathcal{C}|}{S}, \tag{1}$$

where $|\mathcal{C}|$ denotes the number of pixels of the i-th image $I^{(i)}[m,n]$ such that $(m\Delta x, n\Delta y) \in \mathcal{C}$, while S is the overall number of pixel of image i. Hence, the CSA $\alpha(i,j)$ is formally defined as the ratio between the number of pixels belonging to the common areas of two images i and j and the overall number of pixels of the image captured by camera i.

The definition of $\alpha(i,j)$ in (1) allows us to identify the factors that affect the similarity between camera views, accounting for occlusions and uncovered background phenomena between different cameras. The similarity depends not only on the angular distance between cameras but also on the depth and position of the different objects with respect to the camera planes and on the actual occlusions or other acquisition factors. Besides, objects position, depth and relative occlusion vary with time even for steady cameras; being frame-related, the CSA can track the actual similarity better than a geometry-only based parameter.

In the following, we present an empirical evaluation of the MVC efficiency with respect to AVC as a function of the CSA between views, thus providing the rationale for dynamic adoption of the MVC coding scheme in a WMSN. For concreteness sake, in the simulations we approximate the CSA $\alpha(i,j)$ with its estimated value $\hat{\alpha}(i,j)$, computed by means of a low-complexity correlation-based estimator.

The CSA is estimated on a per-frame basis as the number of pixels in the rectangular overlapping region between the view $I^{(j)}[m,n]$, and a suitably displaced version of the second

[1]For the sake of simplicity, we disregard the effect of discretization of the acquisition grid. Nevertheless, such a simplified model can be properly extended to take into account the discrete nature of the acquired image.

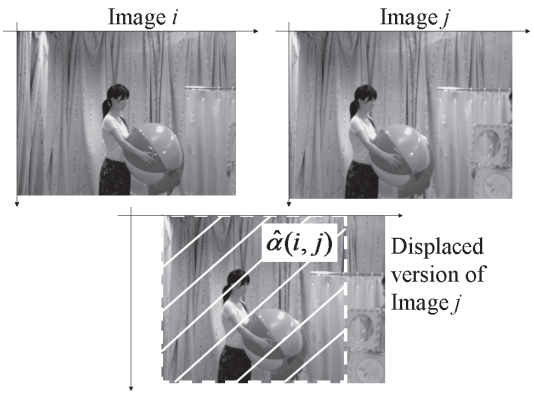


Fig. 3. Example of estimated CSA $\hat{\alpha}(i,j)$.

view $I^{(j)}[m,n]$, as shown for instance in Fig. 3.[2] Despite the coarseness of this computation, this fast estimation techniques allows us to capture the inter-view similarity for the purpose of estimating the MVC efficiency. This approach does not prevent further refinements in CSA estimation, using advanced similarity measures, such as advanced feature-mapping techniques [9]. or even resorting to more refined (but computationally more expensive) view-similarity estimators such as those recently discussed in [10].

IV. Common Sensed Area and MVC efficiency

To quantify the benefits of MVC with respect to multiple AVC, we introduce here the *MVC relative efficiency parameter* $\eta(i,j)$, which depends on the bit rates generated by the codec for the video sequence j when the sequence i is also observed (and therefore known at the encoder).

Let us consider a pair of cameras i and j and let us denote by $r_{AVC}(i)$ the overall bit rate generated by the codec in case of independent encoding (AVC) of the sequence acquired by i-th camera, referred to in the following as i-th view; besides, let $r_{MVC}(i,j)$ denote the overall bit rate generated by the codec in case of joint encoding (MVC) of i-th and j-th views.

The efficiency $\eta(i,j)$ is defined as

$$\eta(i,j) = 1 - \frac{r_{MVC}(i,j)}{r_{AVC}(i) + r_{AVC}(j)}, \qquad (2)$$

and can be interpreted as the gain achievable by jointly encoding the sequences i and j with respect to separate encoding. In case of a pair of sequences, we can also denote as $\Delta r_{MVC}(j;i)$ the bit rate of the differential bit stream generated to encode the j-th view once the bit stream of the i-th view is known, i.e. $r_{MVC}(i,j) = r_{AVC}(i) + \Delta r_{MVC}(j;i)$.

[2]For a given couple of frames, the displacement is chosen so as to maximize the inter-view normalized cross-correlation $\rho(i,j)$, defined as

$$\rho(i,j) = \frac{\sum_{m,n} I^{(i)}[m,n] \cdot I^{(j)}[m+m_0, n+n_0]}{\sigma_i \cdot \sigma_j}$$

with $\sigma_i^2 = \sum_{m,n}(I^{(i)}[m,n] - \sum_{k,l} I^{(i)}[k,l])^2$ and $\sigma_j^2 = \sum_{m,n}(I^{(j)}[m,n] - \sum_{k,l} I^{(j)}[k,l])^2$.

The bit rate generated by the MVC codec depends on the intrinsic sequence activity, which depends on the presence of moving objects in the framed video scene, as well as on the CSA between the considered camera views; for MVC to be more efficient than AVC, it must hold $\Delta r_{MVC}(j;i) \leq r_{AVC}(j)$. In the following, we will show that, as expected, this occurs when the CSA between cameras increases, whereas $\eta(i,j)$ rapidly decays as the CSA between camera decreases. Specifically, we will assess the behavior of $\eta(i,j)$ as a function of $\hat{\alpha}(i,j)$ through experimental tests.

A. Experimental setup with H.264

In order to measure the single-view and multi-view encoding costs, in the simulation we employ the recently defined H.264 MVC [11]; the herein presented study can be straightforwardly extended for different, computationally efficient encoders [2] explicitly designed for WMSNs.

The first video coding experiments presented here have been conducted on a typical MPEG multiview test sequence, namely Akko&Kayo [12], acquired by 100 cameras organized in a 5×20 matrix structure, with 5 cm horizontal spacing and 20 cm vertical spacing. The Akko&Kayo sequence present several interesting characteristics, since the FoVs of the cameras include different still and moving objects (a curtain, persons, balls, boxes), and movements and occlusions occur to different extent. The sequence has been resampled at QCIF spatial resolution, and at 15 frames per second, since such format is compatible with the resourse constrained application framework of WMSN. The experimental results reported here have been obtained using a subset of 6 (out of 100) camera views, i.e., views 0, 5, 10, 20, 40 and 80; the 0-th, 5-th and 10-th cameras are horizontally displaced in the grid while the 20-th, 40-th and 80-th cameras are vertically displaced with respect to the 0-th camera. The first frames corresponding to each of the selected cameras are shown in Fig. 4.

The sequence has been encoded both by means of AVC encoding and by MVC, using a Group of Picture (GOP) structure[3] of $M = 8$ frames; in this latter case, the view of the camera $\#0$ has been selected as the reference view. For fair comparison of the coding results, the MVC coding cost $r_{MVC}(i,j)$ and the multiple AVC coding cost $r_{AVC}(i) + r_{AVC}(j)$ have been compared under the constraint of equal average decoded sequence quality, measured in terms of Peak Signal-to-Noise Ratio[4] (PSNR), Specifically, the PSNR_{AVC}, averaged on the 6 considered views, equals 32.24 dB with 1 dB standard deviation, whereas the PSNR_{MVC}, averaged

[3]In MVC, in the reference view a picture every M is encoded using an INTRA coding mode for the reference view; in the dependent view a picture every M is encoded exploiting inter-view prediction from the contemporary reference view intra frame, using the anchor frame coding mode.

[4]For an $M \times N$ original image $I[m,n]$ represented with l-bit luminance depth, and the corresponding encoded and decoded image $\tilde{I}[m,n]$, the PSNR is defined as

$$\text{PSNR} = \frac{M \cdot N \cdot 2^{2l}}{\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left(I[m,n] - \tilde{I}[m,n] \right)^2}$$

(a) View 0 (b) View 5 (c) View 10 (d) View 20 (e) View 40 (f) View 80

Fig. 4. Selected camera views of Akko&Kayo sequences, horizontal displacement (0-5-10) and vertical displacement (20-40-80).

on the 6 considered views, takes on the value of 32.59 dB with 0.34 dB standard deviation.

B. Experimental Results

First, we present a few simulations remarking that CSA is correlated with the observed MVC efficiency. In Figs. 5 and 6 we plot the behavior of $\eta(i,j)$ for the Akko&Kayo sequence as a function of the frame index for both the horizontal pairs and the vertical ones; for comparison, in Fig. 5 we also plot the MVC efficiency η_{00} of sequence 0 MVC-encoded using as a reference the sequence 0 itself. As already observed in [13] under a different experimental setting, the MVC efficiency achieves its maximum value on frames which are encoded without motion compensation with respect to preceding frames; such frames, realizing both random access and error resilience functionalities, are named *intra frames* in the reference view and *anchor frames* in the dependent views. Besides, the MVC efficiency decreases mainly in the horizontal direction (pairs 0-5 and 0-10) rather than in the vertical one (pairs 0-10, 0-20 and 0-40), and it changes in time (apart from the 0-0 case) because of movement of real objects.

Let us now extend the efficiency analysis to the Kendo multiview test video sequences [12]. For this sequence we have considered 7 views, as acquired by uniformly separated cameras deployed on a line, with 5 cm horizontal spacing. The sequences have been resampled at QCIF spatial resolution, at 15 frames per second. Besides, the different views have been AVC encoded and MVC encoded using view 0 as reference view. As can be observed from Fig. 7 the efficiency values are high for high values of the CSA and rapidly decrease with a CSA reduction. The dotted line represents the common trend of all measured results.

V. Clustering through CSA

To apply the experimental results of the previous Section in a WMSN we consider the following scenario: a set of N randomly distributed sensor nodes equipped with video cameras that can directly communicate to a sink having the task of collecting their camera views. Depending on the network topology and camera orientations, neighboring nodes may acquire overlapping portions of the same scene, leading to correlated views. In our analysis we randomly generated the α values associated to the CSA of each node with its neighbors. Noteworthily, this CSA depends on the proximities of the nodes and also on the overlapping of resulting FoVs.

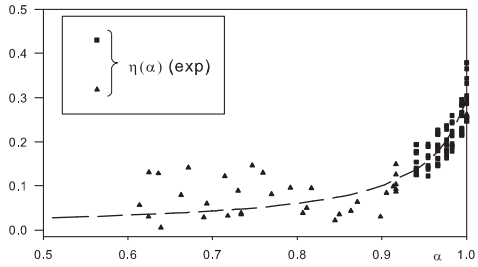


Fig. 7. Scatter plots of the efficiency $\eta(i,j)$ vs estimated CSA $\hat{\alpha}(i,j)$, as observed in several GOPs for the different views in the sequences Akko&Kayo (triangle) and Kendo (square).

The random generation of α models the fact that even close nodes may have different FoVs due to the camera orientations.

We then consider a clustered topology, with m randomly selected cluster heads. Each node can send to the sink video encoded either in AVC or in MVC mode. In the AVC mode, the i-th node generates video at a rate denoted as $r_{AVC}(i)$, i.e., the bit rate for the single-view encoding of the scene acquired by camera i. In the MVC mode the j-th node generates a rate $r_{MVC}(i,j)$ depending on the CSA with the i-th node. In this analysis we derived the $r_{MVC}(i,j)$ according to (2), where $\eta(i,j)$ is derived by the trend reported in Fig. 7.

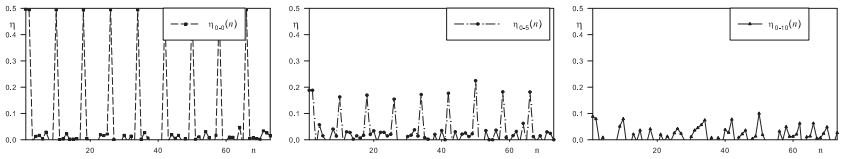
The role of the cluster head is to enable nodes in the cluster to encode their views in MVC mode. To this aim, the cluster-head broadcasts its video in single-view mode. The other nodes use this view as a reference view and generate and transmit their MVC accordingly.

To form the clusters we then define the following scheme:

1) each of the m nodes sends to all its neighbors an image, denoted as thumbnail T_i with $1 \leq i \leq m$, used to derive the $\hat{\alpha}(i,j)$ values associated neighboring nodes [5];
2) each node j receiving the thumbnail T_i computes the $\hat{\alpha}(i,j)$;
3) each node j selects one cluster-head in accordance with the following criteria:
   - $\hat{\alpha}(i,j) < \alpha_{th}$, do not select node i as cluster-head;
   - $\hat{\alpha}(i,j) \geq \alpha_{th}$ select as cluster-head the node i with $\max_{1 \leq i \leq m} T_i$

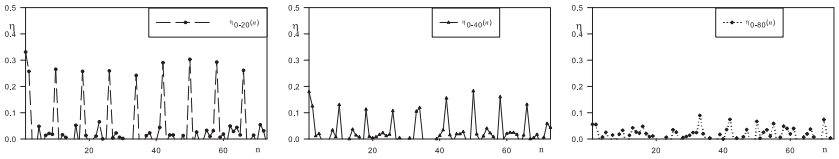
where α_{th} is a threshold set in our correlation model. We consider a thumbnail of size 22×18. As a reference, if a signaling interval of T s is used, transmission of uncompressed

[5]Thumbnails are images at a low resolution entailing low bandwidth and allowing the $\hat{\alpha}(i,j)$ computation



(a) Views 0, 0 (b) Views 0, 5 (c) Views 0, 10

Fig. 5. MVC efficiency as a function of time, using view 0 as a reference: sequence Akko&Kayo, pairs 0-0, 0-5 and 0-10.



(a) Views 0, 20 (b) Views 0, 40 (c) Views 0, 80

Fig. 6. MVC efficiency as a function of time, using view 0 as a reference: sequence Akko&Kayo, pairs 0-20, 0-40 and 0-80.

	Conf1	Conf2	Conf3
Mean efficiency	0.21	0.22	0.20
Max efficiency	0.33	0.40	0.33
Min efficiency	0.10	0.10	0.13
Number of clusters with 1 node	2	2	1

TABLE I
CLUSTER EFFICIENCY IN DIFFERENT SIMULATIONS

thumbnails data would require a bandwidth overhead of $22 \times 18 \times l/T\ bit/s$, being l the luminance depth. For $T = 10s$ and $l = 8\ bits$ this corresponds to an overhead of $316\ bit/s$.

By assuming that all nodes produce the same AVC rate, denoted as R_{AVC}, the overall rate of the cluster i to send all k views of nodes belonging to it to the the sink is $R_{TOT_{MVC}i} = R_{AVC} + R_{AVC} \sum_{j=1,j\neq i}^{k} (1 - 2\eta(i,j))$. It can be observed that this rate depends on the $\eta(i,j)$ that are directly related to the $\alpha(j,i)$. On the contrary, the total rate in the case of single-view transmissions in the i-th cluster is $R_{TOT_{AVC}i} = \sum_{i=j}^{k} R_{AVC_j} = k \cdot R_{AVC}$.

A. Performance in case of clustering

First, we test the gain of the multi-view technology in three different network configurations: Conf1 with $N = 50$ and $R^{AVC} = 80\ kbit/s$; Conf2 with $N = 70$ and $R^{AVC} = 80\ kbit/s$; Conf3 with $N = 90$ and $R^{AVC} = 120\ kbit/s$. In each network we set $m = 10$. Each cluster includes at most N/m nodes and at minimum 1. In some cases the cluster-head was not selected by any neighbors because of the lack of an acceptable correlation between the views ($\alpha < \alpha_{th}$). We set an $\alpha_{th} = 0.5$ and measured the efficiency determined as $1 - R_{TOT_{MVC}i}/R_{TOT_{AVC}i}$ and the mean, min and max values are reported in Table I. It is possible to appreciate that the maximum efficiency achieves high values ($33\% - 40\%$). This indicates that significant advantages can be foreseen in using this approach in clustering schemes.

As a second set of experiments, we generate random networks of different sizes ($N = 2, ...70$) and impose that the cluster size be lower or equal than 3. This is to test if the MVC advantage remains even if the cluster size is kept low, e.g. when the complexity of the cluster (formation and processing) should not increase. We consider a $\alpha_{th} = 0.5$. Fig. 8 shows the MVC and AVC total rate emitted by the network nodes as a function of the network size. The sum of the total rate of each cluster increases as the number of nodes increases and the AVC curve is always above the MVC curve.

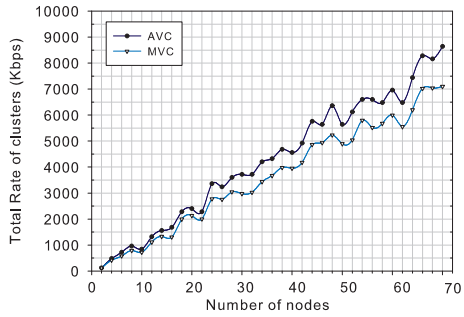


Fig. 8. Total AVC and MVC rates in case of $R^{AVC} = 120kbit/s$ and $\alpha_{th} = 0.5$.

The rate performance gain offered by MVC can also be observed in Fig. 9, which represents the total load of a network of 50 nodes where α_{th} is varied. In this case the number of cluster-heads is assumed to be 10. The number of clusters varies and we observe that the total bit rate generated with MVC is always lower than with AVC but it increases as the α_{th} increases. Therefore, there is a tradeoff in selecting the α_{th}: a high threshold allows producing a low overall cluster bit rate. However, at the same time a $\alpha_{th} \simeq 1$ leads to the isolation of cluster-heads since they cannot find any feasible connections. Consequently, the rate is the same as in AVC since clusters are constituted by only one member.

Then, we assess the role of the cluster size in the considered clustering scheme. We focus on a single network of 50 nodes

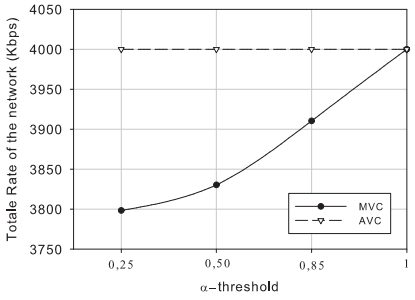


Fig. 9. Total network rate for with different α_{th}; $R_{AVC} = 80\ kbit/s$ and $n = 50$

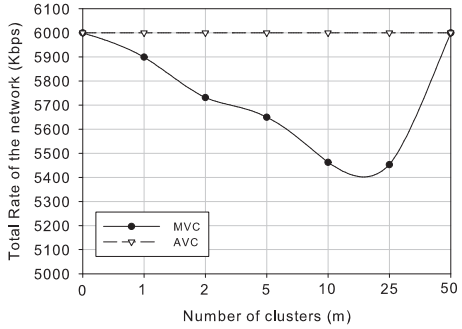


Fig. 10. Total network rate for different cluster sizes; $R_{AVC} = 120\ kbit/s$ and $\alpha_{th} = 0.5$

randomly placed in a given area and with a random selection of the α. We assume a $R_{AVC} = 120kbit/s$ and $\alpha_{th} = 0.5$. Fig. 10 shows how the rate changes as a function of the number of clusters in the network. Obviously, in case of AVC, a clustered size does not affect the rate that is constant and equal to $50 \cdot 120kbit/s = 6000kbit/s$. By adopting MVC instead, we observe that the network load decreases as the number of clusters increases and reaches a minimum for a certain cluster size ($15 \leq m \leq 25$). When the number of cluster-heads increases, the number of single-view coded video increases, since each cluster-head sends its AVC version of the view. Finally, when the number of clusters is equal to the number of nodes ($m = 50$) every node applies single-view coding. From Fig. 9, we observe that the total rate in the network can be optimized as a function of the number of clusters. The optimal value is expected to depend on the sensor spatial distribution as well on the transmission range. These issues are left for future investigations.

VI. Conclusions

We investigated the relationship between the efficiency of Multiview Video Coding and the Common Sensed Area between views. This latter parameter takes into account elements such as the occlusions and the presence of moving objects and thus it captures the common characteristics of view acquired by different nodes giving rise to efficient coding rates that can be applied in Multimedia Wireless Sensor Networks. To show this, we measured the efficiency of H.264 MVC coding and related it with the estimated CSA. Then, we applied the thus found relation in the framework of a WMSN by devising simple intra-cluster coding mechanism. This analysis shows that, by taking into account the CSA, it is possible to enhance the efficiency of the use of the MVC in wireless sensor networks where both the bandwidth and the processing capabilities are reduced. As for the bandwidth we presented several results shoving the benefits of the MVC in the clustering. As for the processing capabilities our tests are based on the estimation of the CSA by using a simplified methodology to capture the inter-view similarity and the application of this methodology on low resolution images (thumbnails) that can be exchanged with few signaling messages in the network.

Acknowledgements

This work has been carried out in the project "Sapienza Social Robot Networks" funded by La Sapienza in 2011. The work of Tommaso Melodia was supported by the US National Science Foundation under grant CNS1117121.

References

[1] R. Dai and I. Akyildiz, "A spatial correlation model for visual information in wireless multimedia sensor networks," *IEEE Transactions on Multimedia*, vol. 11, no. 6, pp. 1148 –1159, Oct. 2009.

[2] S. Pudlewski, T. Melodia, and A. Prasanna, "Compressed-sensing-enabled video streaming for wireless multimedia sensor networks," *Mobile Computing, IEEE Transactions on*, vol. 11, no. 6, pp. 1060–1072, June 2012.

[3] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z.-H. Zhou, "Multi-view video summarization," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 717 –729, Nov. 2010.

[4] A. C. Sankaranarayanan, R. Chellappa, and R. G. Baraniuk, "Distributed sensing and processing for multi-camera networks," *Distributed Video Sensor Networks, Part 2*, pp. 85–101, 2011.

[5] A. R. Vinod Kulathumani, Srikanth Parupati and R. Jillela, "Collaborative face recognition using a network of embedded cameras," *Distributed Video Sensor Networks, Part 5*, pp. 373–387, 2011.

[6] T. Montserrat, J. Civit, O. Escoda, and J.-L. Landabaso, "Depth estimation based on multiview matching with depth/color segmentation and memory efficient Belief Propagation," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, 2009, pp. 2353–2356.

[7] H. Ma and Y. Liu, "Correlation based video processing in video sensor networks," in *Wireless Networks, Communications and Mobile Computing, 2005 International Conference on*, vol. 2, june 2005, pp. 987–992.

[8] P. Wang, R. Dai, and I. Akyildiz, "A spatial correlation-based image compression framework for wireless multimedia sensor networks," *Multimedia, IEEE Transactions on*, vol. 13, no. 2, pp. 388 –401, april 2011.

[9] J.-N. Hwang and V. Gau, "Tracking of multiple objects over camera networks with overlapping and non-overlapping views," in *Distributed Video Sensor Networks*, B. Bhanu, C. V. Ravishankar, A. K. Roy-Chowdhury, H. Aghajan, and D. Terzopoulos, Eds. Springer London, 2011, pp. 103–117.

[10] V. Thirumalai and P. Frossard, "Correlation estimation from compressed images," *J. Vis. Commun. (2012), in press*, 2012.

[11] Y. Chen, Y.-K. Wang, K. Ugur, M. Hannuksela, J. Lainema, and M. Gabbouj, "The emerging MVC standard for 3D video services," in *EURASIP Journal on Advances in Signal Processing*, vol. 2009.

[12] [Online]. Available: http://www.tanimoto.nuee.nagoya-u.ac.jp/ fukushima/mpegftv/

[13] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 11, pp. 1461 –1473, Nov. 2007.