Hierarchical object recognition system in machine vision

A N Kokoulin\(^1\), A A Yuzhakov\(^1\), D A Kiryanov\(^2\), N Stanisavljevic\(^3\) and B Batinic\(^3\)

\(^1\)Perm National Research Polytechnic University, 29, Komsomolsky Avenue, Perm, 614000, Russia
\(^2\)Federal Scientific Center for Medical and Preventive Health Risk Management, 82, Monastyrskaya street, Perm, 614015, Russia
\(^3\)University of Novi Sad, Trg Dositeja Obradovica 6. 21000 Novi Sad, Serbia

Abstract. Authors propose an efficient architecture of the hierarchical machine vision system and the multimedia transfer algorithm for a distributed client-server system which can adapt to mobile network speed variations. This algorithm is based on the wavelet transform of an input multimedia source and allows for data exchange using all the available bandwidth of the unstable mobile network. The main principle of the hierarchical recognition system is the distributed processing network utilizing the “coarse-fine” paradigm in computer vision. Each source of a video stream is processed in-place by the tiny SoC computer which acts as an Edge Computing Unit and detects the presence of an object fragment in a video frame and crops the bounds of the ROI. The resulting stream containing ROI is directed to the main server if the object is detected, reducing traffic and resource consumption.

1. Introduction
Modern information systems utilize the client-server model of data processing. Some of them deal with the multimedia data transfer over the mobile networks and experience difficulties due to the significant download and upload speed variation. The well-known fact is that in spite of the 3G and 4G networks launch in the 2000’s with their outstanding speed, the real speed occasionally degrades in some circumstances such as weather conditions and network equipment utilization. Another fact is the distance effect (Table 1) when the Round Trip Time (RTT) increases [6].

| Distance                  | Network RTT | Avg Packet Loss, % | Throughput, Mbps |
|---------------------------|-------------|--------------------|-----------------|
| Local: < 100 miles.       | 1.6 ms      | 0.6                | 44              |
| Regional: 500–1,000 miles | 16 ms       | 0.7                | 4               |
| One continent: ~3,000 miles | 48 ms   | 1.0                | 1               |
| Two continents: ~6,000 miles | 96 ms  | 1.4                | 0.4             |

So it is not correct to rely on the theoretical values during the information system design and the proper solution is required to utilize all the available network speed at each moment.

This problem can be solved in four ways:
• the better compression of video streams to reduce traffic;
• the Web-optimized storage;
• the use of the decentralized computing scheme with object detection on the edge nodes or intellectual cameras, alignment and cropping of the region of interest (ROI) and ROI transmitting to the server for further processing [1–3];
• the use of the adaptive algorithm of video and image transfer based on wavelet transform.

1.1 The Better Compression
The traffic reduces can be achieved by varying the compression factor of video codec. When the compression is set to 50%, the traffic intensity is reduced up to 7 times and up to 14 times with 100% compression. But at the same time, the efficiency of further processing, for example the object recognition is significantly degraded: starting with the recognition error rate of 99% for the image in the original quality, ending with 55% for the same image with maximal compression factor, (see some experimental results [5] in Table 2).

| Compression ratio, % | Object recognition results, % |
|----------------------|--------------------------------|
| 0                    | 99.5                           |
| 25                   | 95.7                           |
| 50                   | 94.1                           |
| 100                  | 55.2                           |

1.2 Web-optimized Storage

![Web-optimized storage structure](image)

Figure 1. Web-optimized storage structure

Some web services can estimate the current network speed or user’s demands to switch between the different image sources in the Web-optimized system (figure 1). They usually store the full image pyramid which is redundant and needs additional resources to keep it up-to-date.

1.3 The Decentralized (Edge) Computing

It is reasonable to use the concept of distributed Edge computing to process, normalize and aggregate data at the level of network nodes (industrial controllers) and to provide the GRID or cloud computing models on the higher level to analyze data aggregations in the central computer cluster of the SCADA dispatching system. In some cases, the mentioned schema can be successfully ported to satisfy our goals of multimedia data transfer for example if the final aim of the system is object detection and recognition.

The main principle of the hierarchical recognition system is the distributed processing network utilizing the “coarse-fine” paradigm in computer vision [2, 6]: the Object Detection (OD) and the Object Enhancement (OE) stages of object recognition are completed by the “Smart Camera” to determine the
ROI and the Object Recognition (OR) stage is completed on the main server using sophisticated CNN algorithms and centralized databases (figure 2). Each source of a video stream is processed in-place by the tiny SoC computer which acts as an Edge Computing Unit and detects the presence of an object fragment in a video frame and crops the bounds of the ROI.

The resulting stream, containing ROI is directed to the main server if the object is detected reducing traffic and resource consumption. But also we can note that in some cases of occasional mobile network degradation this algorithm fails too, or it can not be suitable if we have to transfer the full-sized image or video.

![Six ROI in the frame](image)

| 2095 Kb full image | 56+63+38+22+49+29=257 Kb ROIs |
|--------------------|-------------------------------|
| Traffic decrease gain 87.73% |

**Figure 2.** ROI detection and processing in distributed computing scheme

The resolution of the ROI image area with the object is lower when the camera resolution is lower. However, for accurate identification of the object, the image in high resolution is needed. The hierarchical method solves this problem.

![ROI resolution and quality enhancement with the hierarchical approach](image)

**Figure 3.** ROI resolution and quality enhancement with the hierarchical approach

Since the transmitted video traffic from all the cameras is significantly reduced, it is possible to use the camera with a higher resolution in comparison with the traditional approach. Thus, the resulting areas with the ROI will have better resolution.

As it is illustrated in figure 3 for the case of the person recognition (identification) system, if we stream the video for the centralized processing in HD-quality, the ROI size can be several times higher.
than if we stream in VGA-quality, and the ROI contains a better-detailed object to make accurate
detection and recognition.

Edge computing allows the cascaded coarse image processing to be executed directly at the smart
camera using HD or even better source video resolution. The result of the OD stage is a high-quality
ROI suitable for OR algorithms.

1.4 The Adaptive Algorithm Of Video And Image Transfer Based On Wavelet Transform
Our previous papers were devoted to the hierarchical storage systems for Big Imagery Data analysis [3, 6] and we continue our research with the hierarchical data processing concept application to the
Computer Vision purposes. Main goals of this project are:

• development and research of the reversible transformation of video source data into a the
hierarchical multi-scale structure with the preservation of all information characteristics. In our previous
theoretical research on this subject, we have shown that the hierarchical representation of data allows
one to significantly reduce the number of I/O operations needed to perform various kinds of user’s
requests and to reduce the consumption of memory and computational resources;

• creation and research of the data storage and processing system model (SDDS) which implements
a hierarchical representation of data at the hardware and software level. We decided to use IoT endpoints
for the primary data collection and object detection processing at the lower level and the main processing
nodes level for object recognition implementation. And we have to adapt the hardware and file systems
to a new format and logical structure of the data being stored and processed at the software site.

The adaptive algorithm of video and image transfer uses the wavelet transform and is based on the
previously developed approach of the reversible transformation of video source data into a hierarchical
multi-scale structure with preservation of all information characteristics.

2. The Reversible Transformation of Source Imagery Data

2.1 JPEG Progressive Encoding
This algorithm was inspired by an interlaced progressive JPEG (and JPEG2000) file format, in which
the imagery data is compressed in multiple passes of a progressively higher detail. This is ideal for large
images that could be displayed while downloading over a slow connection, allowing a low-quality
preview of the whole scene after receiving only a small portion of the data. However, the support for
progressive JPEGs is not universal. To demonstrate the difference between the common and progressive
JPEG compression, see figure 4, where the interrupted transfer of the original JPEG image (left) leads
to different results: the common JPEG image (in the middle) loses the scene totally but the progressive
JPEG (right) preserves the scene and slightly loses some details.

![Figure 4. Standard and Progressive JPEG images with partial download viewing results](image-url)

Using this encoding we can stream images with a constant rate by parts. Each image is divided into
several blocks which are sent one by one by the mobile network until the next image is obtained and is
ready for transfer. The server application assembles the image from the parts which succeeded and we are sure that the scene is transferred to the server. The assembled image has the same quality as the original image if each part was uploaded to the server, and has the lower quality or detalization if any remaining parts were not uploaded. This simple approach is the easiest way to design the transfer method which can adapt to the network speed variation. But there is a problem of the image block size selection which should correspond to the JPEG section size and the trimming of some JPEG file data at the end of the image.

So we decided to design our own image transformation approach based on the powerful and multiscale method of wavelet transform [3] and the video codec utilizing this approach.

2.2 Image and Video Decomposition Using DWT
Advantages of the Wavelet transform compared with Furie transform:
- Multiresolution: Wavelet atoms compress and dilate to analyze imagery data at a wide set of scales. This allows the DWT to match both the short-period and the long-period duration image structures;
- Compression: The wavelet transform results of actual images are known to be sparse and the wavelet coefficient distribution of processed images can be compressed.

The decomposed image is the set of image blocks of several levels, computed using DWT. Basic blocks correspond to the top of the pyramid and we can assemble the image in the lowest resolution of them (figure 5). Values for basic blocks data are aggregated or interpolated by the color values of the neighborhood area. Next level blocks use the basic blocks data as the initial values, which forms the palette. We can form the image of some quality using blocks of both basic and first levels. We can continue assembling the pyramid until we reach the original resolution of the image file, but image data is not considered as the redundant multiresolution stacked representation of the image. Data-driven algorithm of imagery data decomposition and distribution makes it possible to acquire the imagery resources this way: `<img src="http://.../picture.png?width=640&height=480&colors=16"/>` parameters are included in the request and the user receives only necessary data to assemble the image.

We use one data source to serve all requests with different parameters in comparison with redundant Image Pyramid methods.

![Image Decomposition and Assembling](http://.../picture.png)

**Figure 5.** Image decomposition and assembling.

The same data transform approach is used to process the video (figure 6), the same video source is used to stream to all kinds of client applications according to their parameters.
3. The Video CODEC Design

The developed video codec removes the redundancy of the original stream in three ways:

- the color redundancy: removed by converting the RGB color space of the source stream to 16-bit color space;
- the spatial redundancy: removed using a 3-level 2-dimensional Wavelet transform;
- the interframe redundancy: removed by estimating interframe difference.

All stages of the transformations in the pipeline are reversible (figure 7). The codec is based on PackMan - Lossless Video Codec, but the difference is in the adaptive blocked transfer algorithm.

![Figure 6. Video source decomposition and assembling.](image)

![Figure 7. ROI resolution and quality enhancement with the hierarchical approach](image)

The original PackMan codec performs the Reversible Color Transform of RGB 6:6:6 into YCbCr using the following equation:

\[
Y = \text{floor}((R + (2 \times G) + B)/4); \quad Cb = B - G; \quad Cr = R - G
\]

(1)

The reverse transform equation is

\[
G = Y - \text{floor}((Cb + Cr)/4); \quad R = Cr + G; \quad B = Cb + G.
\]

(2)

Our modification is Reversible Color Transform of RGB 8:8:8 into Gr24 and Gr16 (Lossy), explained in figure 8.

![Figure 8. Lossless RGB to Grayscale 24-bit transform](image)

This conversion is close to YCrCb conversion used in practice:

\[
Y = 0.299R + 0.587G + 0.114B
\]

(3)

This conversion is fully reversible and we use single Int32 value containing all the color information.
preserved. If we use the Int16, then the lowest bits of color channels are lost, but the compression is better. Also, we use the Wavelet transform without floating-point multiplications for biorthogonal wavelets to decrease the processor load [4] and entropy Huffman encoding.

This codec uses picture frames of two types: I and P. They differ in the following characteristics:

- I-frames are the most informative and less compressible and do not require any frames to decode.
- P-frames can use data from previous frames to decompress

After the wavelet transform is performed and the H/L byte sequences are calculated and compressed, they are split into blocks of 3 levels to provide the adapting ability. Basic blocks correspond to the lowest quality video frames and are transferred obligatorily to represent the scene with low details. If there is enough time to stream blocks of the second level, they are sent to improve the scene quality. The third level blocks are sent to make the frame quality close to the original source frame. Thus the video streaming algorithm adapts to the current network speed and helps to avoid scene freezing or scene distortion with artifacts.

4. Conclusion

In this paper, the research results of the hierarchical architecture in the distributed object recognition system were proposed. The main principle of the proposed recognition system is the distributed hierarchical processing network utilizing the “coarse-fine” paradigm. Each source video stream is processed by an Edge Computing Unit and detects the presence of an object fragment in a video frame and crops the bounds of the ROI. The resulting stream including ROI is switched to the main server and the object recognition is performed. The traffic decrease gain up to 80-90% is estimated. It is possible to enhance the ROI quality in comparison with the traditional object recognition system, and to improve the recognition accuracy.

The second approach is the adaptive algorithm of video and image transfer based on the wavelet transform for the reversible transformation of video source data into a hierarchical multi-scale structure with preservation of all information characteristics. Using this encoding we can stream images or video with a constant rate by parts. Each image is divided into several blocks which are sent one by one by the mobile network until the next image or videoframe is obtained and is ready for transfer. Thus, the video stream quality is adopted in accordance with the available bandwidth in runtime. We can stream only basic blocks when the channel degrades to preserve the picture on the client site.

5. Acknowledgments

The reported study was partially supported by the Government of Permsky Krai, research project №C-26/174.6 of 31.01.2019.

References

[1] Lowe D G 2004 Distinctive image features from scale-invariant keypoints IJCV 60(2) 91–110
[2] Fleuret F and Gemand D 2001 Coarse-to-Fine Face Detection Int. J. of Computer Vision 41(1/2) 85–107
[3] Kokoulin A 2013 Methods for large image distributed processing and storage IEEE EuroCon 2013 66251911606–1610 DOI: 10.1109/EUROCON.2013.6625191
[4] Kim H 2004 A parallel algorithm for the biorthogonal wavelet transform without multiplication CIS'04 Proc. of the First int. conf. on Computational and Information Science pp 207-212 (Springer-Verlag Berlin, Heidelberg)
[5] Kulikov A A, Demkin D V and Melkov A E 2014 The Analysis of the Maximum Image Compression Impact on Face Recognition Result Science Prospects [in Russian – Perspektivy nauki] 3(54) 104-108
[6] Kokoulin A N 2016 Distributed storage approach in content delivery networks 13th Int. Scientific-Technical Conf. on Actual Problems of Electronic Instrument Engineering, APEIE 2016 – Proc. 2 7806398 pp 479-484 DOI: 10.1109/APEIE.2016.7806398