Analysis of CNN Models to develop a New Appearance Model for Multiple-Object-Tracking

Francisco Javier Alvarez-Prieto
Visual Perception Lab
Tecnologico Nacional de Mexico/I.T.
Chihuahua, Chih, Mexico
fjalvarezp@itchihuahua.edu.mx

Mario I. Chacon-Murguia
Visual Perception Lab
Tecnologico Nacional de Mexico/I.T.
Chihuahua, Chih, Mexico
mchacon@ieee.org

Juan A. Ramirez-Quintana
Visual Perception Lab
Tecnologico Nacional de Mexico/I.T.
Chihuahua, Chih, Mexico
jaramirez@itchihuahua.edu.mx

Abstract— Multiple-object-tracking, MOT, is a paramount research area in computer vision tasks such as: intelligent surveillance systems, motion analysis, self-driving, etc. Different techniques have been considered for the solution of MOT, however, recent works are aimed to involve deep neural networks. Therefore, this paper presents an analysis of different convolutional neural networks (CNN) reported in the literature for MOT. The aim of this analysis is to determine which CNN models perform better when they are evaluated in the main challenges found in MOT. Then, based on the analysis results, a new robust appearance model generated from a Siamese CNN features is proposed. Findings indicate that CNN with fewer layers perform better than CNN with more layers, which may be to the fact that more trivial features allow a better discrimination among the objects of the same class.

Keywords— multiple-object-tracking, appearance model, convolutional neural networks.

I. INTRODUCTION

Video surveillance, sport analysis, event detection, are examples of visual applications that require the intervention of experts to continuously review and analyze recorded videos. This kind of human analysis conveys several disadvantages because of the reduced human attention capacity to perform a repetitive task during a long period of time. In some cases, current technology is used to replace human intervention in these types of activities. However, considering the level of complexity found in this activity more research is still needed. An important factor in the performance of this technology is to have a robust algorithm to achieve multiple-object-tracking (MOT). At the same time, the MOT algorithm relies on its performance in two components: a motion model to predict the position of the objects or targets in the next frames, and an appearance model (AM) to model the aspect of the objects during the video sequence and to establish the correct identity correlation of the different objects. The AM is of great importance, since several authors agree that it is crucial in the performance of a MOT algorithm [1]-[4]. Considering the previous information, this work presents the analysis of several convolutional neural networks (CNN) that may be employed to develop an AM. The analysis, among other issues, considers the ability of the CNN to adapt the AM in different challenges commonly present during MOT.

One contribution of this work is the analysis of several CNN models used to yield AMs, considering their robustness to discriminate objects during the tracking process in several motion challenges; and to propose a new AM based on a Siamese CNN. The experiments show that the Siamese CNN has better discrimination of objects in comparison with the traditional CNN analyzed.

II. LITERATURE REVIEW

MOT is a highly active research area that involves a broad work on the literature. Recently, several AM based on Deep neural networks (DNN) have been proposed in the literature. For example, Hong et al. [5] combine a pretrained CNN with a support vector machine to learn the appearance of the objects by obtaining discriminants from the feature maps. According to their results, the precision increased in the location of the objects with regard the best ten methods in the VT Benchmark. Zhu et al. [6] trained a CNN with spatial attention able to relate coincident patterns of specific sections of two input images, which allow obtaining better performances in terms of metrics to preserve the object identity in comparison with other methods in MOT Benchmark. Yoon and Bae [7] and Kiertitz et al. [8] propose a deep learning method to generate a sequential discriminative AM of each object. This model can track multiple objects online. In [9] Kim et al. describe a new Bilinear Long Short-Term Memory (LSTM) model network. This model allows the sequential long-term learning of objects appearance. Sadeghian et al. [10] combine three recurrent neural networks (RNN) to compute, appearance, motion and social interaction, based on similarity. This combination allows correcting errors in data association and to recover objects identity during occlusions. Fang et al. [11] employ people reidentification techniques and localization information in an autoregressive RNN (RAN). According to their results, the RAN presented better robustness during occlusions in crowd scenes compared to other methods in MOT Benchmark. In other work [12] Chu et al. developed an end-to-end model. In this model, feature extraction, affinity measure and the multidimensional assignment are achieved with only one deep network termed FAMNet. This network is jointly optimized to generate the trajectories of the targets. Considering the results reported in that work, the network acquires the capacity to learn in an integral form the features of high-level affinity. The work of Leal et al. [13] presents a Siamese CNN to relate detected targets between adjacent frames using the output of the Siamese network as a similarity metric. The results of this work indicate a greater accuracy in data association considering other methods in MOT Benchmark. It is important to note that our proposed Siamese Network is different from Leal et al. [13]. Our proposed network has a different architecture. It was defined first by analysis of the deep of other CNN, and its parameters were adjusted considering a discrimination level.

Regarding MOT evaluation methods, the most used in the literature is CLEAR MOT, which considers the metrics MOTP (Multi-Object Tracking Precision) and MOTA (Multi-Object Tracking Accuracy). However, in Section E of this work, a new metric called discrimination level is introduced. This new metric evaluates the performance of AM and can be...
objects of interest are only persons. Fig. 1 illustrates an
database is ideal to test surveillance systems because the
15 frames, taken from the database MOT16 [15]. This
MOT: changes of appearance, scale, illumination, similar
appearance, noise and occlusions [14]. Each category is
described below.

- Appearance change (AC): These scenarios include
objects that change their physical aspects, e.g. objects
seen from different angles.
- Scale variations (SV): This category includes objects
captured by moving cameras, where in the first frames
the objects are small, but as the camera approaches the
object size increases.
- Illumination change (IC): This situation was simulated
by gradually adjusting the brightness of the objects.
- Similar appearance (SA): The set involves objects with
similar physical aspects, e.g. people wearing clothes of
the same color.
- Image noise (IN): The noise was simulated by adding
different levels of Gaussian and salt and pepper noise
to the video.
- Occlusion (OCL): This set includes sequences of
frames where the objects are completely visible at the
beginning of the sequence, and then they are gradually
occluded by other objects until they are completely
occluded.

Each set of test scenarios has five targets captured during
15 frames, taken from the database MOT16 [15]. This
database is ideal to test surveillance systems because the
objects of interest are only persons. Fig. 1 illustrates an
example of each category of the scenarios.

Fig. 1. An example of objects of each one of the 6 categories.

A. Test Scenarios
In order to evaluate the performance of the distinct CNN,
there were considered six types of test scenarios which include
different challenges or categories commonly included in
MOT: changes of appearance, scale, illumination, similar
appearance, noise and occlusions [14]. Each category is
described below.

- Appearance change (AC): These scenarios include
objects that change their physical aspects, e.g. objects
seen from different angles.
- Scale variations (SV): This category includes objects
captured by moving cameras, where in the first frames
the objects are small, but as the camera approaches the
object size increases.
- Illumination change (IC): This situation was simulated
by gradually adjusting the brightness of the objects.
- Similar appearance (SA): The set involves objects with
similar physical aspects, e.g. people wearing clothes of
the same color.
- Image noise (IN): The noise was simulated by adding
different levels of Gaussian and salt and pepper noise
to the video.
- Occlusion (OCL): This set includes sequences of
frames where the objects are completely visible at the
beginning of the sequence, and then they are gradually
occluded by other objects until they are completed
occluded.

B. Selection of the CNN models
The CNN considered in this work are: Alexnet, ResNet-
18, VGG-16, Inception-ResNet-v2, DenseNet-201 and
NASNet-Large. These models were selected to cover the
broad spectrum of CNN strategies proposed in the literature.
Some of the most popular CNN are illustrated in Fig. 2. The
positions of the CNN in the plot correspond to the relation
between their accuracy and their processing time to produce
an output. The CNN selected for this works are indicated in
orange diamonds.

A CNN architecture is mainly composed of two stages:
feature extraction and classification. The first stage includes
convolutional and pooling layers. The second stage is
distinguished by incorporating at least one fully connected
layer to perform classification.

Considering \( F_M \) like the \( i \)th feature map of dimension \( h' \times w' \) in the layer \( l \), where \( h' \) and \( w' \) are the height and width,
respectively. The output of the different layers is computed
as follows.

### Convolutional layer

\[
\begin{align*}
FM_j^l &= f\left( \sum_{i=1}^{i'} FM_{i-1}^l \ast w_j^l + b_j^l \right) \\
FM_j^l &= \text{max}_p\left( FM_{i-1}^l \right)
\end{align*}
\]

where \( b_j^l \) is the bias factor and \( k' \) is the number of filters
used in the layer \( l \), such that \( FM_j^l \) corresponds to the input
image \( O \) when \( l=1 \).

### Pooling layer

This layer reduces the dimension of \( FM_j^l \) by performing a subsampling with a scale factor of \( r \). The
max-pooling technique is the most used subsampling
operation. Thus, a new feature map \( FM_j^l \) is computed by

\[
FM_j^l = \text{max}_p\left( FM_{i-1}^l \right)
\]

here, \( \text{max}_p(\cdot) \) consists on taking the maximum value in
each region \( r \times r \) of \( FM_j^l \).

### Fully connected layer

In the case of the first fully connected
layer, it is necessary to concatenate all features maps of the
previous layer, that is \( DF_{f-1} = [FM_{f-1}^1, \ldots, FM_{f-1}^{k'}] \) to
obtain a vector representation. Then, \( DF_f \) is multiplied with
the weight matrix \( W \) of the neurons in the fully connected
layer. The result of the multiplication is passed through a
non-linear activation function, typically \( \text{softmax} \). Finally, the new
output vector \( DF \) is computed by

\[
DF = f\left( W'\text{DF}_{f-1} + b' \right)
\]

If there are more fully connected layers, the output of
those layers is obtained by (3) but considering the vector \( DF \)
as the input to the new layer, and the weights matrix and bias
of the next fully connected layer.
C. Feature Extraction

To evaluate the performance of the selected CNN features in MOT an experiment was designed. The first stage of the experiment consists on the feature vector extraction of the objects of the six test categories. The second stage evaluates the features of the objects to determine their similarity. The third stage employs the similarity obtained in the second stage to determine the level of discrimination and robustness of the CNN under analysis.

In this work the feature vector, termed descriptor, is obtained by the concatenation of all the feature maps of the layer $L-1$

$$DF^{L-1} = [FM_{j=1}^{L-1}, ..., FM_{k}^{L-1}]$$

(4)

where $L$ indicates the first fully connected layer of the CNN model, Fig. 3. For simplicity and considering that in this experiment each descriptor is obtained in the layer $L-1$, this notation is omitted and $DF^j$ stands for the descriptor of the $ith$ object in the $jth$ frame.

![Image](310x464 to 547x680)

Fig. 3. Obtention of the similarity metric $Ms{/i}$ between objects.

D. Similarity Function

The descriptor of each object is evaluated by a similarity function $As(\cdot)$, to determine the level of similarity $Ms$, used to decide if the target under analysis is the same target or not. Two $As(\cdot)$ are employed in this work. The cross-correlation coefficient $\rho$, and the Euclidean distance $Ed$. Before the computation of $Ed$, the descriptors $DF$ are normalized to unitary vectors represented by $\overline{DF}$.

$$\rho = \frac{\sum_{n=0}^{N-1} (DF^j_k[n]) \cdot (DF^i_k[n])}{\sqrt{\sum_{n=0}^{N-1} (DF^j_k[n])^2} \cdot \sqrt{\sum_{n=0}^{N-1} (DF^i_k[n])^2}}$$

(5)

$$Ed = \sqrt{\sum_{n=0}^{N-1} (DF^j[k] - DF^i[k])^2}$$

(6)

where $DF^j_k$ is the descriptor of the $kth$ object in the $jth$ frame. Therefore, $Ms{/i}$ represents the similarity, Fig. 3, between the target $Oj$ with respect the target $Oi$. Thus, a value of $Ms{/i}$ close to 1 obtained with the correlation $\rho$ or a value close to 0 using $Ed$ indicates that the targets are more similar.

![Image](48x461 to 287x551)

Fig. 4. Similarity measure using the descriptors of CNN Alexnet (a,b), and Inception-ResNet-v2 (c)(d).

E. Discrimination Level $Dl$

The discrimination level $Dl$ is computed as follows. Assume the similarity metric between the object of interest with itself

$$MsOI^j = As(DFjj, DFjj^{-1})$$

(7)

and the similarity metric of the object of interest and the other objects as

$$MsOP^j = As(DFjj, DFjj^{-1})$$

(8)

Note that only the similarity metric is computed on consecutive frames ($j$ and $j+1$). Therefore, the appearance model is an updated appearance model because it is updated after processing each video frame. Thus, the separation margin in the frame $j$ for the object $i$, $Sei^j$, is given by (9) when $\rho$ is employed, and for (10) when $Ed$ is used.

$$Sei^j = MsOI^j - \max(MsOP^j)$$

(9)

$$Sei^j = \min(MsOP^j) - MsOI^j$$

(10)

Now, since each test scenario contains five objects captured during 15 frames, the $Dl$ of each CNN for a specific challenge is computed by the average of the separation margin considering the 15 frames for the five objects (11).

$$Dl = \frac{1}{5} \sum_{i=1}^{5} \frac{1}{14} \sum_{j=1}^{14} Sei^j$$

(11)

$Dl$ can be interpreted as the easiness to distinguish an object of interest from the other objects in the test, in a specific
MOT challenge. So that, as higher the value of \( DL \) the challenge will be overcome. However, it is important to note that there is no direct comparison between the results using \( \rho \) and \( Ed \) as the similarity function.

Table I shows the values of \( DL \) obtained for each CNN in each one of the six test scenarios when \( \rho \) is employed, and Table II for the case of the \( Ed \). This information is also illustrated in Fig. 5 and Fig 6 respectively.

**TABLE I. \( DL \) of the CNN using \( \rho \) and the updated AM approach.**

<table>
<thead>
<tr>
<th>CNN</th>
<th>IC</th>
<th>SA</th>
<th>AC</th>
<th>SV</th>
<th>OCL</th>
<th>IN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexnet</td>
<td>0.62</td>
<td>0.36</td>
<td>0.46</td>
<td>0.39</td>
<td>0.45</td>
<td>0.33</td>
</tr>
<tr>
<td>VGG-16</td>
<td>0.72</td>
<td>0.40</td>
<td>0.49</td>
<td>0.41</td>
<td>0.46</td>
<td>0.38</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>0.17</td>
<td>0.13</td>
<td>0.13</td>
<td>0.14</td>
<td>0.18</td>
<td>0.11</td>
</tr>
<tr>
<td>DenseNet-201</td>
<td>0.37</td>
<td>0.21</td>
<td>0.26</td>
<td>0.26</td>
<td>0.35</td>
<td>0.24</td>
</tr>
<tr>
<td>Inception-ResNet-v2</td>
<td>0.16</td>
<td>0.09</td>
<td>0.12</td>
<td>0.09</td>
<td>0.16</td>
<td>0.11</td>
</tr>
<tr>
<td>NASNet-Large</td>
<td>0.34</td>
<td>0.20</td>
<td>0.23</td>
<td>0.20</td>
<td>0.30</td>
<td>0.26</td>
</tr>
</tbody>
</table>

**TABLE II. \( DL \) of the CNN using Ed and the updated AM approach.**

<table>
<thead>
<tr>
<th>CNN</th>
<th>IC</th>
<th>SA</th>
<th>AC</th>
<th>SV</th>
<th>OCL</th>
<th>IN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexnet</td>
<td>1.02</td>
<td>0.53</td>
<td>0.60</td>
<td>0.46</td>
<td>0.50</td>
<td>0.43</td>
</tr>
<tr>
<td>VGG-16</td>
<td>1.06</td>
<td>0.46</td>
<td>0.55</td>
<td>0.41</td>
<td>0.44</td>
<td>0.38</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>0.51</td>
<td>0.28</td>
<td>0.30</td>
<td>0.26</td>
<td>0.32</td>
<td>0.21</td>
</tr>
<tr>
<td>DenseNet-201</td>
<td>0.74</td>
<td>0.33</td>
<td>0.42</td>
<td>0.35</td>
<td>0.43</td>
<td>0.35</td>
</tr>
<tr>
<td>Inception-ResNet-v2</td>
<td>0.49</td>
<td>0.22</td>
<td>0.28</td>
<td>0.20</td>
<td>0.29</td>
<td>0.24</td>
</tr>
<tr>
<td>NASNet-Large</td>
<td>0.71</td>
<td>0.31</td>
<td>0.38</td>
<td>0.28</td>
<td>0.37</td>
<td>0.37</td>
</tr>
</tbody>
</table>

**Fig. 5. \( DL \) of the CNN using \( \rho \) and the updated AM approach.**

**Fig. 6. \( DL \) of the CNN using Ed and the updated AM approach.**

At this point, the experiment only considered the similarity between objects in consecutive frames. In order to verify the robustness of the AM built from the descriptors of the CNN, a new experiment using a non-updated AM was achieved. That is, the AM is generated in the first frame and it is kept without change during the complete video sequence. In this experiment the similarity metric of the object of interest with itself, and the object of interest with the other objects in the test are computed by

\[
MsO\ell_i = \lambda s(D_i^\ell, D_{i+1}^\ell) 
\]

and

\[
MsO\ell_i = \lambda s(D_i^\ell, D_{i+1}^\ell) 
\]

In this experiment, the similarity metric is computed between the objects in the frame 1 and the frames \( j+1: j=1, 2, \ldots, 14 \). This allows to test the robustness of the descriptors because the similarity metric is computed temporally with more distant objects, which may include drastic changes. The \( DL \) in this experiment is also computed by (9)-(11). The \( DL \) obtained for each CNN using \( \rho \) in this version of the experiment is shown in Table III and Fig. 7, and Table IV and Fig. 8 show the values of \( DL \) employing \( Ed \).

**TABLE III. \( DL \) of the CNN using \( \rho \) and non updated AM approach.**

<table>
<thead>
<tr>
<th>CNN</th>
<th>IC</th>
<th>SA</th>
<th>AC</th>
<th>SV</th>
<th>OCL</th>
<th>IN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexnet</td>
<td>0.59</td>
<td>0.24</td>
<td>0.23</td>
<td>0.17</td>
<td>0.19</td>
<td>0.27</td>
</tr>
<tr>
<td>VGG-16</td>
<td>0.67</td>
<td>0.25</td>
<td>0.23</td>
<td>0.13</td>
<td>0.16</td>
<td>0.23</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>0.19</td>
<td>0.13</td>
<td>0.07</td>
<td>0.07</td>
<td>0.08</td>
<td>0.04</td>
</tr>
<tr>
<td>DenseNet-201</td>
<td>0.35</td>
<td>0.13</td>
<td>0.14</td>
<td>0.14</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td>Inception-ResNet-v2</td>
<td>0.24</td>
<td>0.07</td>
<td>0.09</td>
<td>0.03</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>NASNet-Large</td>
<td>0.31</td>
<td>0.10</td>
<td>0.11</td>
<td>0.08</td>
<td>0.13</td>
<td>0.16</td>
</tr>
</tbody>
</table>

**TABLE IV. \( DL \) of the CNN using Ed and non updated AM approach.**

<table>
<thead>
<tr>
<th>CNN</th>
<th>IC</th>
<th>SA</th>
<th>AC</th>
<th>SV</th>
<th>OCL</th>
<th>IN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexnet</td>
<td>0.86</td>
<td>0.31</td>
<td>0.27</td>
<td>0.18</td>
<td>0.20</td>
<td>0.27</td>
</tr>
<tr>
<td>VGG-16</td>
<td>0.87</td>
<td>0.27</td>
<td>0.23</td>
<td>0.12</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>0.46</td>
<td>0.21</td>
<td>0.14</td>
<td>0.10</td>
<td>0.12</td>
<td>0.05</td>
</tr>
<tr>
<td>DenseNet-201</td>
<td>0.62</td>
<td>0.19</td>
<td>0.20</td>
<td>0.16</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>Inception-ResNet-v2</td>
<td>0.52</td>
<td>0.13</td>
<td>0.16</td>
<td>0.04</td>
<td>0.08</td>
<td>0.10</td>
</tr>
<tr>
<td>NASNet-Large</td>
<td>0.56</td>
<td>0.16</td>
<td>0.16</td>
<td>0.10</td>
<td>0.15</td>
<td>0.18</td>
</tr>
</tbody>
</table>

**Fig. 7. \( DL \) of the CNN using \( \rho \) and the non updated AM approach.**

**Fig. 8. \( DL \) of the CNN using Ed and the non updated AM approach.**

The results generated when an update AM is used indicate that Alexnet and VGG-16 report the best performances. That means that in these experiments Alexnet and VGG-16 are
more recommended solving the different MOT challenges. It is worth mentioning that Alexnet and VGG-16 are the least deep CNN considered in the experiment. This may suggest that as the deep of the CNN decreases its capacity to differentiate objects increases. This issue is different from other CNN applications like classification, where the deepest CNN performs the best.

The previous results can be explained by looking at the information encoded by the deepest layers of the CNN. As mentioned in [16], the information in the deepest layers is associated with the semantic content of the objects and as the depth of the CNN increases, the quality of semantic information improves. Semantic information is particularly useful to discriminate objects from different classes, but not when objects belong to the same class. Therefore, networks with less capacity to describe semantic content (those that are shallower) obtain better results to discriminate objects of the same class.

In the case when the AM is not updated the less deep CNN are still the best CNN. However, their performance is considerably reduced compared with an AM updated approach. This may lead to conclude that none of the CNN studied are robust enough to model drastic changes of object appearance.

### F. Proposed Appearance Model

This section describes the proposed AM based on a Siamese CNN configuration. This configuration involves two identically CNN with the same parameters. The main difference between this neural architecture and a traditional CNN is that the Siamese network is fed with two images, that may be a positive pair (images of the same object) or a negative pair (images of different objects). Both images are propagated through the network to generate their respective descriptors, and then the similarity between them is computed. Thus, the output of the network is the similarity value. One advantage of this network configuration is that the similarity and dissimilarity concepts are directly related to the learning algorithm of the network.

The architecture of the Siamese network was defined by considering the previous experiments and the depth of the CNN analyzed. It was decided to design a network with a 50% reduction in the convolutional layers compared to CNN Alexnet, which is the least deep CNN considered in the experiment. Thus, the Siamese network has three convolutional layers (C_l), three pooling layers (P_l) and a fully connected layer (FC) represented by equations (1) to (3). Fig. 9 illustrates the proposed Siamese network, where the final block corresponds to the similarity function \( \rho \) defined in (5). The information of each layer is presented in Table V.

The training of the Siamese network was achieved by batches of size of 180 and 1500 iterations. The learning rate was 0.00001 and the cost function was the binary cross-entropy. The training images were from the database CUHK03 [17], which was designed for person reidentification task in views of 10 different angles. Equal number of positive and negative pairs were provided during the training. The images were normalized by the z-score normalization process. To produce image values with zero mean and unitary standard deviation.

<table>
<thead>
<tr>
<th>Layer type</th>
<th>Layer properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image input</td>
<td>128x64x3 with z-score normalization</td>
</tr>
<tr>
<td>Convolutional</td>
<td>12 filters 15x15x3 with stride [1 1], padding [0 0 0 0]</td>
</tr>
<tr>
<td>layer 1, C_1</td>
<td>and ReLU</td>
</tr>
<tr>
<td>Max pooling</td>
<td>2x2 max pooling with stride [2 2] and padding [0 0 0 0]</td>
</tr>
<tr>
<td>layer 1, P_1</td>
<td></td>
</tr>
<tr>
<td>Convolutional</td>
<td>24 filters 9x9x12 with stride [1 1], padding [0 0 0 0]</td>
</tr>
<tr>
<td>layer 2, C_2</td>
<td>and ReLU</td>
</tr>
<tr>
<td>Max pooling</td>
<td>2x2 max pooling with stride [2 2] and padding [0 0 0 0]</td>
</tr>
<tr>
<td>layer 2, P_2</td>
<td></td>
</tr>
<tr>
<td>Convolutional</td>
<td>48 filters 3x3x24 with stride [1 1], padding [0 0 0 0]</td>
</tr>
<tr>
<td>layer 3, C_3</td>
<td>and ReLU</td>
</tr>
<tr>
<td>Max pooling</td>
<td>2x2 max pooling with stride [2 2] and padding [0 0 0 0]</td>
</tr>
<tr>
<td>layer 3, P_3</td>
<td></td>
</tr>
<tr>
<td>Fully connected</td>
<td>4096 neurons with sigmoidal</td>
</tr>
</tbody>
</table>

To validate the performance of the proposed network, it was compared with the CNN VGG-16 because it obtained the best performance with the similarity function \( \rho \) in the experiment of Dl. Fig. 10 and 11 show the comparison considering the updated and non-updated AM approaches, respectively.

![Fig. 9. Appearance model based on a Siamese CNN.](image)

![Fig. 10. Comparison between the Siamese and VGG-16 networks considering Dl, \( \rho \), and the updated AM approach.](image)

![Fig. 11. Comparison between the Siamese and VGG-16 networks considering Dl, \( \rho \), and the non-updated AM approach.](image)
IV. RESULTS

The results illustrated in Fig. 10 show that the proposed model outperforms the results of the VGG-16 in the challenges AC, SV, and IN, and slightly better in IC. VGG-16 is better in the challenges SA and OCL. Nevertheless, it can be noticed in Fig. 11, that the proposed model is far better than VGG-16 in AC, SV and IN, it is slightly better in IC and OCL, and it is only overcome in SA.

An important point is that the DI in the Siamese network stays almost the same in both experiments. On the other hand, the DI of VGG-16 drastically decays when the AM is not updated. This is an indication that the proposed network can model the appearance of objects even they are not in consecutive frames. Therefore, it is a more robust network to face drastic object changes. However, this advantage limits its capacity to distinguish different objects with similar appearance. These results, as it was mentioned in Section E, are related with the semantic content present in the layers of the CNN and the concepts of similarity and dissimilarity introduced during the training of the Siamese network.

V. CONCLUSION

This paper presented an analysis of different CNN used to generate an AM used in MOT. Based on this analysis a new Siamese network was proposed for MOT.

The experimental results indicate that CNN with fewer layers like the Alexnet and the VGG-16 can discriminate better objects in different MOT challenges. Therefore, deeper CNN employed in this work present lower capacity to discriminate objects in MOT.

Regarding the proposed network, the findings show that the Siamese network yielded better results in most of the experiments involving different MOT challenges than the other CNN considered in this work, except in the SA case.

ACKNOWLEDGMENT

This research was funded by Tecnologico Nacional de Mexico/ I.T. Chihuahua under grants 5162.19-P and 7598.20-P.

REFERENCES


